



Visualization and exploration of texts



A theoretical framework and two practical approaches
for improvement of digital libraries and information retrieval systems

PhD Thesis by Jaume Nualart Vilaplana

Department of Information Science
University of Barcelona
MMXV

Visualization and exploration of texts.

**A theoretical framework and two practical approaches for improvement of
digital libraries and information retrieval systems**



UNIVERSITAT DE BARCELONA

**DEPARTAMENT DE BIBLIOTECONOMIA I
DOCUMENTACIÓ**

Doctorat en Informació i Documentació en la Societat del Coneixement

Curs 2014-15

Visualization and exploration of texts

**A theoretical framework and two practical approaches for improvement
of digital libraries and information retrieval systems**

**Tesi doctoral presentada per Jaume Nualart Vilaplana per optar al títol de
doctor per la Universitat de Barcelona**

Director de tesi: Doctor Mario Pérez-Montoro Gutiérrez

Contents

Acknowledgements	5
Abstract	6
Overview	7
1. Introduction	11
1.1. Text visualization	11
1.2. Structuring text visualization	13
1.3. Search engine results and text visualization	14
1.4. Digital Libraries and text visualization	14
2. Literature review	16
2.1. The lack of a classification schema	16
2.2. The lack of visualization of search results	17
2.3. The lack of visualization of digital libraries	18
3. Aims and research questions	20
3.1. Classification for text visualization	21
3.2. Search engine results interfaces	21
3.3. Digital libraries interfaces	22
4. Methodologies	23
4.1. About the corpus	23
4.2. Classification for text visualization	23
4.3. Search results and single text visualization	24
4.4. Digital libraries and text collection visualization	25
5. Results	27
5.1. Classification for text visualization	27
5.2. Search results and single text visualization	34
5.3. Digital libraries and text collections visualization	38
6. Discussion and conclusions	41

Bibliography	43
A. Appendix: Published papers	49
A.1. Paper I	49
A.2. Paper II	66
A.3. Paper III	83
B. Appendix: Forty-nine text visualization approaches	92

Acknowledgements

Jordi Kiami, Kiese Garbí, Amelia.

Maria Rosa i Jaume i la resta de família catalana i angolana.

Als germanets de vida Stephane, Ivan, Juan, Toni, Montalvo, Joan i Josep Lluís Peris.

A la Gabriela per ser amiga nostra, per ser com és i per l'ajut i suport continuus. A la Joelle per l'ajut, l'amistat i la coincidència. Al Giulio, per les converses més interessants possibles, per la força creadora, la sinceritat, i per ser com és. Al Luke, el meu mate, que ens ha fet sentir com a casa.

Als amics que sempre ajuden Chris, Mar.

Òbviament, al Mario, el millor mentor possible, gràcies per la teva paciència, empena, sinceritat, precisió, complicitat i experiència, i gràcies sobretot per la teva amistat.

Al moviment per al programari i la cultura lliures i a totes les persones que lluiten per un coneixement lliure i sense propietat.

This work is licensed under a Creative Commons Attribution-ShareAlike 3.0 Unported License.

Cover by J.L.Montalvo (montalvo.club) - thanks!

Abstract

In English:

This research project aims to improve the way humans work with textual documents when doing tasks such as exploring, discovering, searching, filtering, collecting, indexing, comparing, or just reading. This research project studies theoretical foundations and practical uses of text visualization techniques. As a contribution to theory, the research project presents a classification schema for text visualization approaches based on visual features instead of task-solving capabilities (Paper I). As a contribution to practice, how to improve interfaces of digital libraries (DL) has been studied. Two practical proposals for approaching text are introduced: one for single text representation, called Texty (Paper II), and another for text collections exploration and overview, called Area (Paper III). This research project discusses the contrast between the growing popularity of text visualization, presented as a subfield of data visualization, and the lack and urgency, nowadays, of interactive and visual interfaces to DL.

En Català:

Aquest projecte d'investigació té com a objectiu millorar la forma en la qual els humans treballem amb documents de text quan fem tasques com explorar, descobrir, cercar, filtrar, compilar, indexar, comparar, o simplement llegir. Aquest projecte de recerca estudia fonaments teòrics i usos pràctics de tècniques de visualització de text. Com a contribució a la teoria es presenta un esquema de classificació per a casos de visualització de textos basat en les característiques visuals de cada cas en lloc de en llur capacitat de solucionar tasques concretes (Article I). Com a contribució a la pràctica s'ha estudiat com millorar les interfícies de les biblioteques digitals (DL). S'introdueixen dues propostes pràctiques per a abordar textos: una per a representar textos individuals, anomenat Texty (Article II) i un altre per a explorar i donar una visió de conjunt de col·leccions de textos, anomenat Àrea (Article III). Es discuteix el contrast entre la creixent popularitat de la visualització de textos, presentat com un subcamp de la visualització de dades, i la manca i urgència, avui en dia, d'interfícies interactives i visuals per a biblioteques digitals.

Overview

In 2015, data is one of the top trending words. Never before has such an amount of data been produced, “yet a lot of the richest information we have is in text format” (Heer, 2010). The study of texts has fascinated humans for a long time, but today it seems urgent to advance the development of tools to help deal with the huge amount of data to which an increasing number of people are exposed to in everyday life.

These two reasons – fascination and urgency – drive the beginning, the evolution and the results of this research project and this dissertation.

This thesis is presented as a compendium of publications. It includes a dissertation, which presents the research project. This dissertation is followed by the three publications that the research has produced so far. The publications are referred to as Paper I, II, and III. Full text of the papers can be found in Appendix A.

This research project is a study of text visualization, its theoretical foundations, and its practical use. As a contribution to theory, I present a classification schema for text visualization approaches (Paper I). As a contribution to practice, I have studied how to improve interfaces of Digital Libraries (DL). Most DL use text-based interfaces, which have not evolved a lot over the last two decades, especially when compared to, e.g., e-commerce websites. Following the proposed classification schema of text visualization approaches, I have studied the basic features of DL, and what DL websites offer to their visitors. I developed two practical proposals for approaching text: one for single text and another for text collections (papers II and III, respectively). The two solutions proposed are complementary, i.e., they could be applied simultaneously, and they try to answer different research questions (see Section 3. Aims and research questions).

This dissertation starts with an Introduction where I present the conceptual structure of the research project as a consistent unity. In the introduction, text visualization is also presented as a new field within data visualization. Section 2 is a literature review, which introduces Section 3, aims and research questions guiding the project. In Section 4, I present the methodologies used in this research project to answer the questions raised. In Section 5, the longest section, I present all the results produced by the research. Finally, I discuss the results and present the conclusions. Appendix A contains three papers. Appendix B presents the forty-nine reviewed cases of text visualization approaches.

Paper I:

Nualart-Vilaplana, Jaume; Pérez-Montoro, Mario; Whitelaw, Mitchell (2014). How we draw texts: a review of approaches to text visualization and exploration. *El profesional de la información*, mayo-junio, v. 23, n. 3, pp. 221-235. <http://dx.doi.org/10.3145/epi.2014.may.02>}

Abstract This paper presents a review of approaches to text visualization and exploration. Text visualization and exploration, we argue, constitute a subfield of data visualization, and are fuelled by the advances being made in text analysis research and by the growing amount of accessible data in text format. We propose an original classification for a total of 49 cases based on the visual features of the approaches adopted, identified using an inductive process of analysis. We group the cases (published between 1994 and 2013) in two categories: single-text visualizations and text-collection visualizations, both of which can be explored and compared online.

- Published in: “El profesional de la información”.
- Journal indexed in: ISI Social Sciences Citation Index (Q3), Scopus (Q2), and more.
 - SJR (SCImago Journal Rank) (2014): 0.374
 - Impact factor (JCR), IF (2014): 0.356
- Selection: Editor first selection, and then double-blind peer-reviewed.
- Other: Open Access publication.

Paper II:

Nualart, Jaume; Pérez-Montoro, Mario (2013). Texty, a visualization tool to aid selection of texts from search outputs. *Information Research*, 18(2) paper 581. [Available at <http://InformationR.net/ir/18-2/paper581.html>]

Abstract Introduction. The presentation of the results page in a search system plays an important role in satisfying the information needs of a user. The usual performance management criteria and tools to organise results have limitations that may hinder the satisfaction of those needs. We present Texty as a new approach that can help improve the search experience of users.

Method. The corpus of texts to which we applied Texty were papers from Information Research. To filter the texts, we have built five groups of words or vocabularies on concrete fields of knowledge: conceptual approach, experimental approach, qualitative methodology, quantitative methodology and computers/IT.

Results. We show how Texty, intrinsically, is capable of encoding or offering its users information about the text that other alternative classic representations (bar or lines charts, mainly) are not able to offer.

Conclusions. Texty is a complementary tool that improves intellectual interaction with a list of texts, allowing users to choose texts more effectively knowing their structure before reading them.

- Published in: Information Research
- Journal indexed in: ISI's Web of Knowledge, Scopus (2013 Q2)
 - Also indexed in: INSPEC: Engineering Village, Library, Information Science & Technology Abstracts, LISA: Library and Information Science Abstracts.
 - SJR (SCImago Journal Rank) (2014): 0.254
 - IPP (Impact per Publication) (2014): 0.418
 - SNIP (Source Normalized Impact per Paper) (2014): 0.447
- Selection: Editor first selection, and then double-blind peer-reviewed.
- Other: Open Access publication

Paper III:

Pérez-Montoro, Mario; Nualart, Jaume (2015). Visual articulation of navigation and search systems for digital libraries, *International Journal of Information Management*, Volume 35, Issue 5, October 2015, Pages 572-579, ISSN 0268-4012

<http://dx.doi.org/10.1016/j.ijinfomgt.2015.06.005> <http://www.sciencedirect.com/science/article/pii/S0268401215000614>

Abstract Journal and digital library portals are the information systems that researchers turn to most frequently for undertaking and disseminating their academic work. However, unlike other information systems, such as e-commerce websites, their interfaces have not been improved on the basis of the findings provided by user studies, nor have the advances developed in specific disciplines, such as information architecture, or those derived more generally from User Experience (UX), been applied to them. In an attempt at overcoming these limitations, from the late eighties onward, a series of prototypes have been developed that seek to improve the visualization of results from journal and digital library portals. Yet, these prototypes and advances in visualization have not been widely adopted for both practical and methodological reasons. To overcome these limitations, new solutions and low-cost tools that can be readily implemented, and which can improve user-experience and user-satisfaction with these information systems, need to be identified. One possible

alternative is the articulation of the navigation and search systems in a single visual solution that would allow the simultaneous exploration and interrogation of the information system. Area is a new, low-cost visualization tool that is easy to implement, and which can be used with large collections of documents. Moreover, it has a short 1 learning curve that articulates the two systems using a two-dimensional structure which enhances both user-experience and user-satisfaction with journal and digital library websites.

- Published in: International Journal of Information Management (IJIM). Elsevier.
- Journal indexed in: Scopus, and Web of Science (Categories: Information Science & Library Science).
 - Also indexed in: CMCI, Communication Abstracts, Computer & Control Abstracts, Computer Literature Index, Contents Pages in Management, Current Contents/Social & Behavioral Sciences, Current Technology Index, International Political Science Abstracts, Library and Information Science Abstracts, DBLP, PAIS Bulletin, Research into Higher Education Abstracts, SSCI, Sociological Abstracts.
 - Quartiles by category in 2014: Computer Networks and Communications (Q1), Information Systems (Q2), Library and Information Sciences (Q1).
 - SJR (SCImago Journal Rank) (2014): 1.093
 - IPP (Impact per Publication) (2014): 2.820
 - SNIP (Source Normalized Impact per Paper) (2014): 2.111

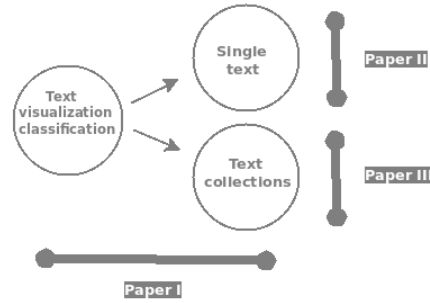
Selection: Editor first selection, and then double-blind peer-reviewed

Other: Elsevier have the copyright of the version published in IJIM

1. Introduction

This chapter introduces the general topic of this research, text visualization, and the three subtopics: classification schema, single text visualization, and text collections visualization. The research has generated three papers on those topics, which have been published in scientific journals (see Figure 1).

Figure 1. Simple diagram of the research



Paper I is a contribution to the theoretical framework of the text visualization field. Papers II, and III are practical applications of the studied features that are not covered in the reviewed literature.

1.1. Text visualization

In 2015, data visualization can be considered as a consolidated new field of knowledge (Strecker and Wind, 2012). Several indicators confirm this. Eight of the top ten universities according to the Times Higher Education ranking 2014-15 (Education Ltd TLS, 2015) have specific departments or research groups working in the field of data visualization. The number and the importance of conferences and journals dedicated primarily to data visualization is growing (see Paper I).

Table 1. Eight out of ten top universities have departments devoted to data visualization (2015)

Institution	Rank 2014- 15	Department/Group	URL
California Institute of Technology (Caltech)	1	<i>Human Interfaces Group</i>	http://www.hi.jpl.nasa.gov/work-with-us/
Harvard University	2	Broad Institute of Harvard and MIT	http://www.broadinstitute.org/vis
University of Oxford	3	Visual Informatics Lab at Oxford	http://oxvii.wordpress.com/
Stanford University	4	Stanford Vis Group	http://vis.stanford.edu/
University of Cambridge	5	re—	—
Massachusetts Institute of Technology	6	Broad Institute of Harvard and MIT	http://www.broadinstitute.org/vis
Princeton University	7	PrincetonVisLab	http://bit.ly/1UXtie3
University of California, Berkeley	8	VisualizationLab	http://vis.berkeley.edu/
Imperial College London	9	Data Science Institute	http://www.imperial.ac.uk/data-science/
Yale University	10	—	—

New studies are presented under the name data science in universities and other educational institutions around the world. Data science has become a buzzword and, as such, has multiple definitions and uses. The overuse of the buzzwords “open data” and “big data” in political campaigns and by governments has also contributed to spreading the interest and increasing centrality of data and its social and power implications in many of today’s societies. Data science has always been a multidisciplinary field. This is shown in figure 2. Data science, as used in this text, comprises a heterogeneous group of disciplines and techniques that, together, specialises in the study of phenomena based on data measures and production. A data scientist must be able to collect, analyse, and communicate evidence found in studying data. Data visualization is evolving from a multidisciplinary nature to a well-delimited brand new discipline.

Figure 2. Word cloud of the professions practiced by inventors of visualization methods from 1765 to 1999 (see Paper I).



Text visualization should be considered as a well-defined field because of its relationship with text analysis. The analysis of texts uses specific methods that belong to the study of textual documents (text similarity, topic models, text summarisation, and information extraction), and not to other data types.

Text visualization can simply be defined as any representation of a text or a collection of texts that is shown in an interface using graphic elements in addition to text. This interface can be interactive or not. Usually, the representation includes an overview of the text or the collection, as defined in Shneiderman's Visual Information-Seeking Mantra: Overview first, zoom and filter, then details-on-demand (1996).

The next three subsections briefly introduce the three convergent contexts of this research: structuring text visualization, search engine results, and digital libraries.

1.2. Structuring text visualization

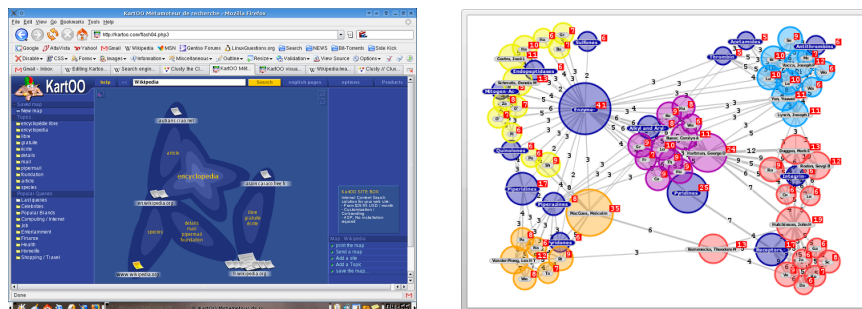
Paper I argues that one of the consequences of the youth of a scientific field is its lack of standards in sources of new knowledge, I.e., conferences and publications, and in vocabulary, methodology, the evaluation of new ideas, and in their classification according to a theoretical framework. Standards spread, get used, get acceptance, and finally, comes a sort of assimilation. This process requires time among other factors. Consequently, text visualization, as a new field, has only a few set standards. A classification schema is a critical tool (Lancaster, 1972). In the study of a knowledge field. An important aim of this project is to fill this gap by proposing a classification framework for text visualization approaches.

In chapter 3 I reference other classifications and justify why I decided to propose a new classification for text visualization cases.

1.3. Search engine results and text visualization

Paper II argues that information retrieval systems typically present the results of a query in flat, one dimension lists. Usually, these lists are opaque in terms of order, I.e., the users do not know why the list has a particular order. In fact, most search engines present results in that way. Two old examples of graphic results are KartOO (extinct) (KartOO, 2009) and TouchGraph (Shapiro, 2003), today a company that offers graphic results only as a paid service. See Figure 3.

Figure 3. Two graphical search engines: KartOO, 20012010 (left) and TouchGraph (right)



Text entries from 2002 show that graphical search result interfaces were a promising technique that never have been massively adopted (Shaphiro, 2002). The non-visual evolution of the presentation of search engine results is surprising. There is room for experimenting with text visualization techniques and for studying why adoption of new visual interfaces and interactions used to fail.

1.4. Digital Libraries and text visualization

Paper III presents an in-depth review of digital information systems, and more particularly, of online DL. A DL is a collection of documents in digital format that can include text, images audio and video, and is designed to satisfy the needs of visitors to locate information within a collection. DL are usually organised and structured using a classification that is based on one or more specific criteria related to the content. A navigation system is defined to allow users to move comfortably around the different categories of the collection. Controlled vocabularies and other artefacts – thesauri, taxonomies, synonym rings – also contribute to the navigation (Pérez-Montoro, 2010). Nevertheless, the two elements most frequently used by users when seeking information are the search and navigation systems. Thus, the search box and the navigation bar are clearly identified in digital information system interfaces.

In order to study these topics, I have selected an online DL as a case study for text visualization. A not so big and accessible DL is a good laboratory to work with. The data is clearly delimited and, in the best of cases, well indexed.

In the next section, I review the related literature to justify my research and highlight some gaps.

2. Literature review

Text visualization can be defined as a sub field of data visualization for factual and strategic reasons. In paper I, I found controversy about this and some authors tend to disagree: Illinski (2013) claims that text cannot be considered a data type; Šilić (2010) argues that “unstructured text is not suitable for visualization”. It is true that most text visualizations transform the initial “unstructured” textual data into a reduced, structured dataset; in which the new dataset is no longer one-dimensional, but rather constitutes a categorical or a network dataset that can be represented with a wide range of tools that are not specific to text representation (Hearst, 2009, Grobelnik and Mladenic, 2002). In Paper II, I demonstrate that it is also possible to visualize text as it is, i.e., as unstructured one-dimensional data. In Paper III, I did not work directly with the texts at all, but with the metadata of each represented document.

Paper I argues that, to study text visualization, it is important to examine the literature dedicated to both data visualization and text analysis, given the significant interrelationships that exist. Thus, while the text analysis output may limit the possibilities of visual presentation and interaction with the text, there is strong empirical evidence indicating that people learn better with a combination of text and graphic elements than with text alone (Anglin et al., 2004, Levie and Lentz, 1982). For this reason, I argue that there are factual and strategic reasons to consider text visualization as a differentiated field of knowledge. I argue that factual reasons are due to the fact that everyday text is the most produced type of data, and that strategic reasons are because such an approach promotes text visualization research projects. Moreover, it would stimulate collaboration among analysis and visualization of texts.

2.1. The lack of a classification schema

A reason for questioning the discussion that text visualization is a field is probably the lack of an established theoretical framework. This project can contribute to filling that gap.

In Paper I, forty-nine cases of text visualization have been reviewed in order to study how other researchers faced the challenges of using visualization techniques in DL interfaces. Once the cases were collected, and the analysis was started, I found it difficult to group the cases. None of the existing classification schemes were appropriate for my research because I wanted to classify cases according to visual features, and existing classifications were task oriented. I reviewed the classical classification schema of Shneiderman (1996) and from Collins et al. (2009), both of which were based on tasks that the visualization approach

can solve rather than on the explicit aspects of the visualization itself. In a classification based on tasks, two different cases that look quite different are good to solve same or similar tasks. I wanted to differentiate cases based on their visual and graphical strategies that can be used in different contexts and for different tasks. Certainly, our classification is related to visual language and diagrammatic understanding instead of problem-solving.

2.2. The lack of visualization of search results

As argued in Paper II, on one hand, most search engines results are represented as a flat list. On the other hand, information retrieval is a critical factor in an environment characterised by excess of information (Baeza-Yates et al., 1999). The presentation of search results plays an important role in satisfying user information needs. It is widely accepted that a bad or inadequate presentation can hinder the satisfaction of the information needs (Shneiderman, 1992, Baeza-Yates et al., 2011, Hearst, 2009, a). As mentioned above, the results of a query as a one dimension list in terms of order are opaque to the users. Paper II analyses many of the proposed alternatives to the simple, flat list of results. It differentiates a query within one document or within a collection of documents. In the first case, the system outputs the document, highlighting those words that literally match the terms of the query (Egan et al., 1989). Some studies indicate that users prefer to see this technique implemented by using colour to match the query terms in highlighted words (Hornbaek and Frokjaer, 2001). In the second case, each document is represented on the results page as a horizontal bar, proportional to extension of the document, and small squares are added for each query term that appears in the text (Hoeber and Yang, 2006). As in the previous case, some studies indicate that this representation improves with the introduction of a colour scale proportional to the frequency of the query terms in the document (Anderson et al., 2002). Another technique for visualizing search results is to add thumbnail images or miniaturized images of the retrieved documents on the page. This technique is based on the fact that the human visual system captures the essentials of an image in 110 milliseconds or even less, just what it takes to read a word or two (Woodruff et al., 2001). Some studies claim that adding these images in the search results transforms them into visual summaries of documents (Jhaveri and Raelihae, 2005).

Several techniques have been applied to search engine results: clusters, treemaps, tag clouds, networks. All these techniques can improve the search experience of users, but they all have important limitations. Compared to visualizations of clustering, the treemaps provide extra information on the thematic focus of the retrieved documents and the semantic relationships among them. However, the treemaps do not provide information on the distribution and thematic structure of each document. Tag clouds also provide extra information on the thematic focus of the retrieved document, but they do not provide information on possible semantic relationships between documents, nor orientation on the distribution and thematic structure of each document. Finally, network graphs provide extra information on

the thematic focus of the retrieved documents and possible semantic relationships between documents, but do not provide the distribution and thematic structure of each document.

Visualizations based on the search query terms also have important limitations. They only provide documents in which the query terms appear. They do not provide extra information on the thematic focus of the retrieved documents, nor possible semantic relationships between retrieved documents. They do not give any orientation on the distribution and structure of the terms that are unrelated to each of these retrieved documents either.

Finally, the visualization strategy which involves completing the list with thumbnail images or miniaturised images of retrieved documents also has important limitations. These visualizations, though complementary, do not provide extra information. Studies show that the thumbnail images strategy does not significantly improve the search experience of users (Czerwinski et al., 1999, Dziadosz and Chandrasekar, 2002), although they can be helpful in part if the images are enlarged (Kaasten et al., 2002).

These limitations lead us to seek new forms of visualization that can help to improve the search experience of users in information retrieval systems and any other case where the user has to choose or select documents from one dimension lists of documents.

In the following section, I explain how I propose to face the challenge of visual search results.

2.3. The lack of visualization of digital libraries

As mentioned in the previous section, a lot of effort has been carried out to introduce the use of visualization techniques in search results of digital information systems. However, when referring to DL, and especially online publications, the reality shows that most DL use classic text-based interfaces. Again, the search box, the navigation bar, and the categories schema are the preferred elements to serve contents to the visitors. This traditional representation has significant limitations. On the one hand, it does not always provide sufficient information about the content of a document to enable the user to accept it or dismiss it without having to read or interact with it first (Baeza-Yates, 2011, b, Nualart-Vilaplana et al., 2014). On the other hand, it does not allow the user to deploy techniques of berry-picking in the search process (Bates, 1989). This could refine the results obtained so as to propose subsequent, more efficient searches, in keeping with the user's changing information needs following interaction with the results.

In Paper III, I present a variety of techniques that have been applied to the exploration and overview of both search engine results and DL : two-dimensional visualizations using maps or clusters (Chalmers and Chitson 1992, Andrews et al. 2001, Anderson et al. 2002), two-dimensional tables or grids (Fox et al. 1993, Shneiderman et al. 2000, Kim et al. 2011), and three-dimensional visualizations (Robertson et al. 1991, Hearst and Karadi 1997, Cugini et al. 2000, Hienert et al. 2012). These visual prototypes make a series of significant improvements to the classical interfaces of scientific journal websites and DL. Thus, on the

one hand, they provide shorter search times compared to traditional non-visual methods (Hienert et al., 2012) and, on the other, they enable a more efficient formulation of queries tailoring the information needs of users. Finally, they provide additional information to users, information that is not available on a page of more conventional results. This extra information, which shows semantic relationships between the documents retrieved, provides a better interaction with the results and facilitates the refinement of subsequent queries (Bauer, 2014).

Yet, even with these significant improvements to the classical interfaces, these prototypes and advances in visualization have not been widely implemented in the portals or websites of journals or DL. As argued in Paper III, the reasons for this are varied and can be classified as practical and methodological reasons.

Practical reasons include that, in resources of this type, these tools are implemented as separate pages from the basic search interfaces, which means that users perceive them as secondary tools. Furthermore, these improvements, especially those that visualize the results, involve a high level of abstraction and conceptualisation, making them not very intuitive for users. Perhaps more importantly, implementing these techniques, unlike traditional interfaces, does not offer any clear commercial or economic benefit in the world of digital systems of scientific information of this type (Van Hoek and Mayr, 2013, van Hoek and Mayr, 2014).

Methodological reasons include that very few of the proposed techniques have been tested and evaluated with end users, which makes it difficult to draw any clear conclusion about their efficiency.

In the next section, I present a proposal to address those challenges.

3. Aims and research questions

The general aim of this research project is to improve the way that textual documents are used, read, understood, and consumed. The project is driven by two observations about the misuse of DL: the way textual documents are explored, and the way search results are represented. That they are misused in the way search results are represented, affects not only DL but also search engines. Most DL, including heritage archives and scientific journals, have a similar interface to interact with. As mentioned in the introduction, other types of web interfaces, e.g., e-commerce websites, have evolved considerably, but DL websites have continued to stay the same.

On the one hand, classic online exploration is based on tree category browsing; usually it lacks a quantitative overview of the collection (number of items, and weight of query terms within the collection), and qualitative representations of the collection (geolocation, timeline, author-documents view, topics analysis view). On the other hand, classic search engine features include a search box for quick search, advanced search form filter, and a flat list of search results. Internally, the system gets better as the research in information retrieval advances, but externally, the only visual evolution of those interfaces is in graphic design, where new styles make forms look more attractive and, eventually, more usable, even when doing same tasks.

I have not investigated why this happens because this task is beyond the scope of this research project. My position at this point has been to work positively, contributing ideas to the evolution of interfaces and the way we interact with text documents and collections of text documents.

From this observation, I can formulate two initial research questions:

- How can the traditional flat list of search engine results using visualization techniques be improved without interfering with information retrieval operations?
- How can collections of documents (DL) be explored and searched using an interface based on graphic elements and visual language instead of only-textual interfaces?

To address these questions, I collected and reviewed cases of text visualization from academic literature and online publications. Once the list of cases to review was set, as explained in previous sections, a third question was developed:

- How can text visualization cases be classified according to visual features?

This order of questions reflects the reasoning process. However, the research steps were started by this last question.

3.1. Classification for text visualization

The classification question, then, is a consequence of the study of cases of text visualization in academic literature and in independent and personal publications. The research study demanded to map text visualization approaches in order to find patterns, similarities and gaps among them, and group them accordingly. The process of classifying manually case by case, and discussing aspects of each case, has brought consistency to the proposed classification schema.

I propose a faceted classification that starts with the question: single text approach or text collection approach? Then, depending on the given answer, more questions related to the visualization and task capabilities follow:

Single text approach:

- Whole text or part of the text?
- Does the visualization follow the same sequence as that of the text?
- Does the visualization use elements from the discourse structure or from syntactic structure of the text?

Text collection approach:

- Are the items of the collection differentiated or represented as aggregations?
- Just data or data and landscape?

For both single texts and collections:

- Is the dataset dynamic or static?
- Is the visualization a result of a search query?
- Is it valid for small or large datasets?

Section 5 Results, presents the classification schema specifications and its performance.

3.2. Search engine results interfaces

The literature shows advantages in using some graphic elements attached to the traditional flat list of search results: thumbnails of the items in the list of search results, and small charts (i.e. sparklines) attached to each item of the list. These techniques are discussed in Subsection 2.2.

To combine these insights, Paper II presents Texty, a single text visualization approach that is not interfering with information retrieval ranking algorithms and complements the traditional flat list of search results attaching an image to each item from the list of results. Each image is icon-like and represents the physical distribution of keywords within a text. These keywords are grouped in five vocabularies, to each of which a colour is linked. Texty reveals the structure, conceptual density and subject matter of each item from a list of results. This could help the users decide which item to explore.

Texty is presented in detail in Section 5 Results.

3.3. Digital libraries interfaces

Paper III proposes a graphic interface to DL that focuses on information performance and usability features. In order to expose and test these features with users, I present an interface called Area. Area uses the same interface to explore, search and filter DL. I developed an early version in Perl in 2007. I used its conceptual base to completely rewrite the software in 2014. It now is a modern only-client-side JavaScript application.

As in Paper II, for this study, I have used the same text corpora: the open access journal Information Research.

The strategy with Area has been to offer an interface with all the features that the Information Research existing interface offers, plus several new visual features. The dataset and this methodology are explained in detail in the next section.

Both interfaces, the existing one and the proposed one, have been compared through a user evaluation study. According to the evaluation, on average, 80% of participants preferred the proposed interface to the existing one. See Section 5 Results for a detailed description of the user study and other outputs.

4. Methodologies

In this section, I present the actions and experiments conducted in order to answer the research questions. I also introduce the corpus used for the experiments with DL and search results outputs.

4.1. About the corpus

The data used in this research project, both to explore and search in DL using an interface based on graphic elements and to improve search results flat list outputs, is a corpus of 592 scientific papers derived from the open access journal Information Research, edited by Prof T.D. Wilson since 1995 until present (Information Research journal, 2015). The papers are published in HTML under the Creative Commons license: Attribution-NonCommercial-NoDerivs 2.0 UK: England & Wales (CC BY-NC-ND 2.0 UK). The description of the journal states:

“Information Research is a freely available, international, scholarly journal, dedicated to making accessible the results of research across a wide range of information-related disciplines. It is privately published by Professor T.D. Wilson, Professor Emeritus of the University of Sheffield, with in-kind support from Lund University Libraries, Lund, Sweden and from the Swedish School of Library and Information Science.”

This corpus accomplishes several conditions that make it a suitable corpus for this experiment. It is a controlled collection of texts with a similar register and a specific semantic field, it is freely accessible and has an open license, which encourages reuse (Wilson, 2013). Moreover, the papers belong to the same document collection, have unity, share the academic register, have a similar structure (introduction, method, analysis, results) and have a standardised quality (peer-reviewed).

4.2. Classification for text visualization

As explained in Paper I, the methodology to collect the cases involves two stages. First, a traditional literature review, including not only academic publications but also practical examples and demos, published on websites of reference – including Infosthetics, Visual-complexity and Visualizingdata.com – and personal blogs. The reason for including non-academic sources is that the visualization community is composed in a similar proportion

by academic researchers and practitioners, from non-profit, freelance studios, and companies. Second, a subset of the collected cases have been selected, based on a preliminary analysis of their features that seek: text related works, originality, novelty, and in general, cases that provided a representative overview of the range of works in the field. The final selection was, therefore, more exclusive than inclusive, and this is more qualitative than quantitative (Benavides et al., 2010).

The classification of the cases is the product of empirical observation following an inductive analysis, that is, the observation and selection of a variety of cases is used to find patterns and propose an initial classification schema. This schema is again tested with the cases and modified again. These iterations produce a final classification schema that embraces all the features found in the reviewed cases in the best-reached way.

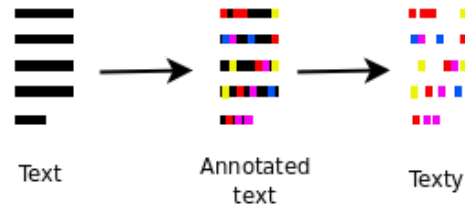
4.3. Search results and single text visualization

The process to create the so-called ‘Texty’ from the text of the Information Research papers is explained step by step in the next paragraph. For an extended description of each step, refer to Paper II.

The steps followed to answer the research questions are:

- Decide which corpus to use. As mentioned, Information Research corpus was selected according to the requirements described in Section 4.1.
- Choose semantic categories. I identified five subject categories that could help to classify the contents of scientific papers: conceptual approach, experimental approach, qualitative methodology, quantitative methodology and computers/IT. The election of the semantic categories, though related to the corpus of texts, is not unique and could be different without affecting the ideas of Texty as a possible helpful tool.
- Define sources for the corpora of the vocabularies. Each category needs to be defined with a vocabulary, i.e., a list of words. To do this, I did choose some concepts for each category and took their definitions from three sources: English Wikipedia, Stanford Encyclopedia of Philosophy, and Britannica Encyclopedia. See Table 1 from Paper II.
- Process terms for each vocabulary. First, a stopwords filter was used to remove empty words. Then, words occurring less than four times were deleted, as they were considered of little significance for each subject. Then, words occurring in more than one vocabulary were deleted, i.e., I removed intersections between vocabularies. Thus, I obtained a number of terms for each vocabulary (Table 2 of Paper II). Finally, terms were reviewed in order to detect terms that were inconsistent with the subject matter, ambiguous or not coherent with each vocabulary.

Figure 4. Simplified diagram of Texty generation process



- Define the final five vocabularies and their number of terms as: Conceptual Approach, 65 terms; Experimental Approach, 53 terms; Qualitative Methodology, 74 terms; Quantitative Methodology, 86 terms; and Computers/IT, 410 terms. The final list of terms can be found online (Jaume Nualart, 2013, a).
- Get the data from Information Research. In order to study the data in the cleanliness of the laboratory, I cloned the Information Research website on a working server. The Textys produced can be found integrated with a copy of the Information Research site dated on the date of the experiment. Find the Texty demo online (Jaume Nualart, 2012, b), access with user “texty”, and password “texty”.
- Parse and annotate the text. For each paper, the words belonging to each of the vocabularies have been annotated using HTML class tags.
- Get the Texty images. The whole paper in HTML has been colour-styled appropriately and converted to images, generating the so-called texty.
- Evaluate the resulting tool. Texty features are compared with bar charts and line charts, showing that Texty gives a more accurate and much deeper information about the text than classic charts.

4.4. Digital libraries and text collection visualization

In order to answer the research question, I developed a practical data visualization project. The project actions and methods applied during this development are grouped in: data gathering, software development, and user evaluation study. Below, I present a list of the actions and methods briefly described above. For more detailed information, see Paper III.

About the data:

- Metadata specification: Fields to collect and their type were defined, and human names for fields were labelled.

- Scrapping Information Research papers: This has been laborious work, mainly because the HTML format for the journal has changed a lot and multiple times since 1995 until today. I mainly worked with the python spider Scrapy, and complemented with python and shell scripting. All the codes created for this research are published under free licenses. See Section 5 for details.

About the software:

A totally new version of the software has been coded following the specifications:

a) General specifications:

- JavaScript only-client-side application.
- Data stored in online JSON files.
- The use of SVG and D3.js to increase visualization performance.

b) Particular specifications:

- Some pre-existing features from the Information Research interface have been used as they are, and simply redirected from the new interface to the existing one.
- Two existing features have been redefined – the lists: by-author and by-subject.
- Some other features are brand new that the proposed interface, Area, brings out of the box. The main menu of the existing interface is also present in the new interface, as well as the logo and journal description, so the usual visitor gets access to the existing website sections and can feel certain familiarity in the new interface.

About the evaluation:

- a) User evaluation: An online survey was undertaken. The goal of this survey was to know whether participants are open to the implementation and use of visual interfaces. To answer this question, I asked the participants how they would solve common tasks when visiting scientific online journals, as well as which of the two interfaces suits each of the proposed tasks better.
- b) Performance evaluation: An example with dummy data allows the performance of the application to be tested according to the number of registers represented and the length of its metadata (see Subsection 5.3 Performance evaluation results).

5. Results

In this chapter, I present a summary of the results. For a more detailed report, see Papers I, II and III. As in most of the chapters, this chapter is presented in three sections, one for each related topic as follows.

5.1. Classification for text visualization

The reviewed cases are included in Appendix B, and are referenced in the text by title, author(s) and ID number (1 to 49 according to appendix numeration). The cases can be explored online using an Area representation. Area is, in turn, the tool introduced in Paper III (Jaume Nualart, 2015c).

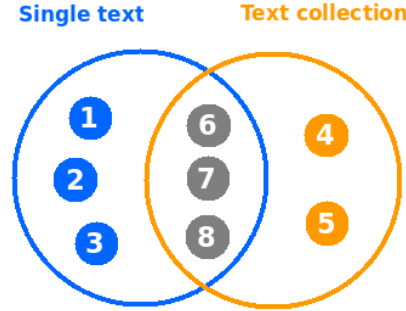
Definitions of the classification schema

The classification is designed according to multiple taxonomies, and it is a faceted classification.

Two groups of facets are defined: single text approaches and text collection approaches. As shown in Figure 5, eight facets were defined: three facets specific to single text approaches, two specific to text collections, and three applicable to both of them.

In practice, as explained in Chapter 4 Methodologies, to classify a text visualization approach, a number of conditioned questions related to the features of the case need to be answered. The first question to answer is what type of corpus is represented? Single text, or texts collection? Depending on the answer, a second round of questions will follow.

Figure 5. Venn diagram of the eight binary questions to classify text visualization approaches



For single texts:

1) Whole text or part of the text? In some cases, a part from the text is considered the essence of the text and is used as the main component of the visualization approach instead of using the whole text.

Nevertheless, there are representations where the whole text participates. These cases often represent the whole text in an implicit way. Examples of this are:

- The chapters of a book as differentiated visual entities.
- The representation of all the sentences of the text as coloured lines.
- The verbs of a text, providing a representation of the style of the text.
- The characters of a novel and their appearance within the text.
- The places or dates present in the text.

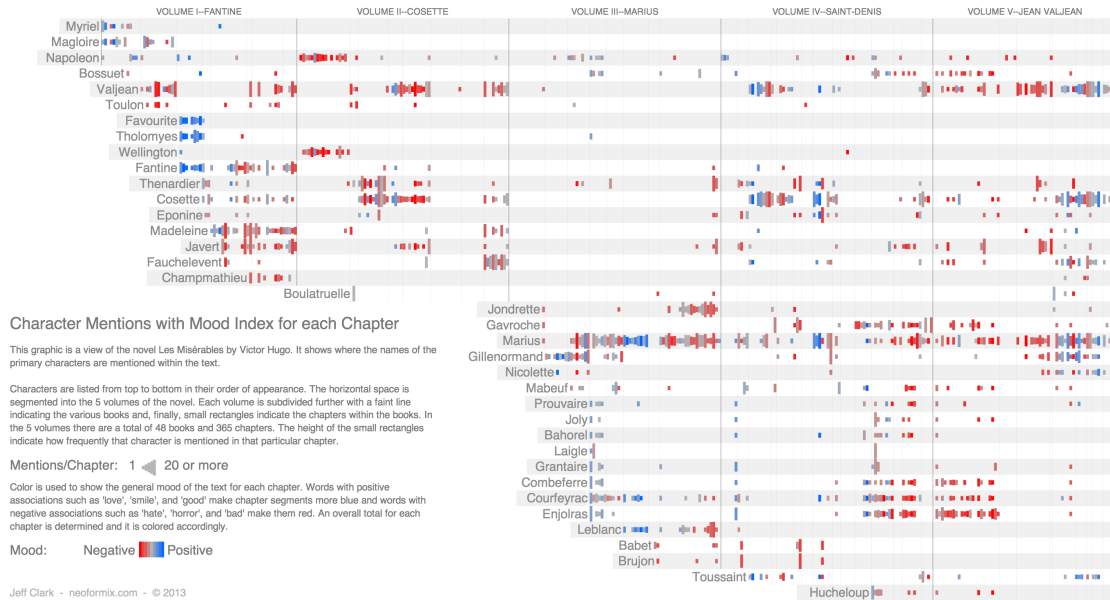
The cases in which the whole text is explicitly represented, for obvious reasons, are cases with short texts, e.g. texts of songs, speeches, and poems, among others.

In some cases, such as case #2 “Novel Views: Character Mentions by Jeff Clark” (See Figure 6) only certain words within the text are represented; nonetheless, I classify this case as a whole text representation because the whole novel, chapter by chapter, is implicitly represented along the circle.

In cases in which the whole text is represented, even implicitly, by a single element of the visualization, I have classified it as whole text visualization.

Figure 6. Novel Views: Les Misérables - Character Mentions by Jeff Clark (2013): the whole novel, chapter by chapter, is implicitly represented along the circle.

NOVEL VIEWS - Les Misérables - Character Mentions



2) Does the visualization follow the same sequence as that of the text? Is the visualization following the same sequence as the original text? By sequence, I mean the same order as in the text. If yes, the case will be considered sequential. Otherwise, it will be called non-sequential.

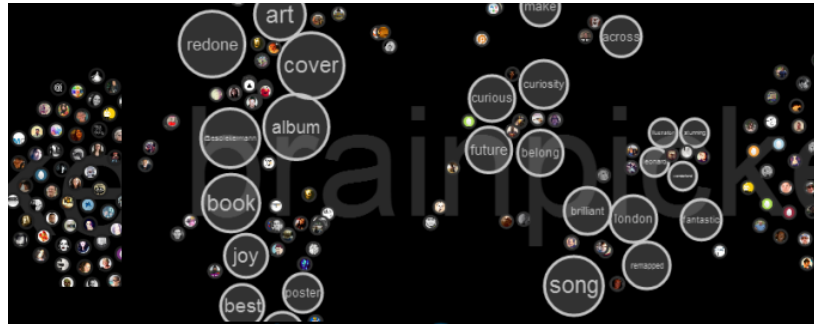
For example, a typical case that does not follow the sequence of the original text is a word cloud (see case #17 “Wordle by Jonathan Feinberg” and see Figure 7).

3) Does the visualization use elements from the discourse structure or from the syntactic structure? Among others, a text can have two kinds of structure that I consider useful for our research. There is a structure that is completely subjective to the author’s point of view, and this is the so-called discourse structure. In linguistics, discourse is a broad concept. Here I am using it to refer to the parts of a text in concordance with the outline of a document; that is: parts, chapters, sections, subsections, etc. The discourse structure is widely used when visualizing texts because it is a direct way to draw a representation of the text sequence.

The second structure that I consider is the syntactic structure. Syntactic can also refer to several concepts. Here I am using it to refer to: sentences, phrases, and words such as verbs, nouns, and morphemes. This is an objective structure and depends on the rules of

Scales and axes are not considered as landscape, nor are the functional elements of the interface (buttons and menus) in which the tool is embedded.

Figure 8. Detail of case #26 Spot by Jeff Clark (2012): tweet topics and tweet authors are represented in two dimensions. The distances between elements are related to their similarity.



For both single texts and collections:

Properties applicable to single text and to texts collections visualizations are:

6) Is the dataset dynamic or static? Are the texts changing over time? There is a set of visualization tools that show the changes of a dataset over time. Most popular tools of this kind have been developed in computer science, representing code evolution, or with Wikipedia data, showing a number of aspects of historical article editions. I also include media visualizations of latest news. Since in the control version systems for coding and in Wikipedia examples, every bit of data is dynamic by definition, In the latest news example, the dataset basically just grows over time. All those cases are included in this category.

7) Is the visualization a result of a search query? The visualization of search engine results is characterised by the changing number of represented items. This number depends on the number of results obtained for a query. This is a visualization subfield related to the disciplines of information systems and information retrieval (Mann, 1999, Hearst 2009).

8) Is it valid for small or large datasets? It is rare that a visualization tool is independent of the size of the data that is represented in it. Most cases are applicable to small, medium or big datasets. The definition of dataset size is heavily dependent on the field and the kind of research project. It is also possible that, in some cases, the data size is not an issue, e.g. a word cloud visualization is independent of the size of the text source. Text analysis carries the performance related to data size.

Quantitative analysis of reviewed cases and results of the classification

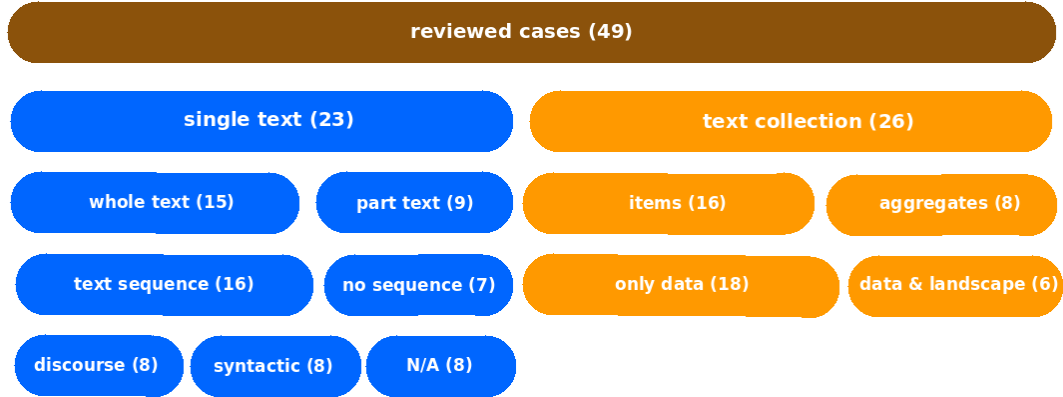
The aims of the review of forty-nine cases is to collect a qualitative representation of ideas and approaches in text visualization, and use them to refine and prove the consistency of the classification schema presented.

The quantitative analysis that follows is key to understanding the proposed classification schema. The classification schema is an endogenous consequence of the reviewed cases. I used a reflective practice methodology: study a case, test it with the set of questions, review the questions, retest the case with the questions.

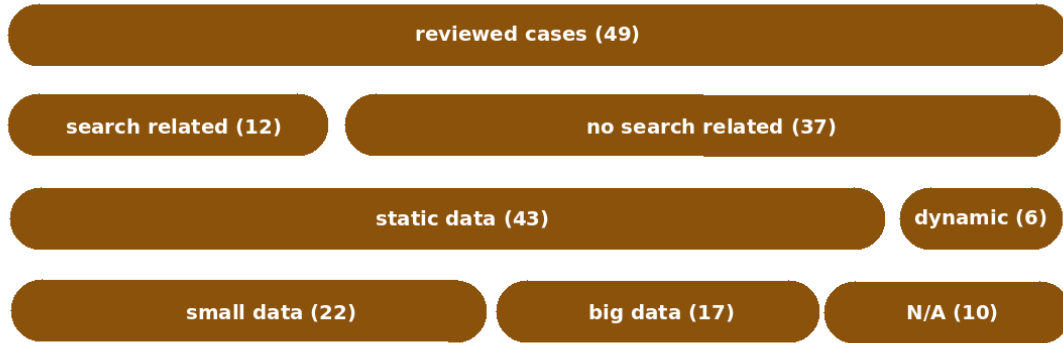
The produced schema is a faceted classification with two groups of questions: for single text, and for text collection visualizations. Each question can have three answers: A, B or N/A (not applicable).

According to the type of data used in each case, I have found twenty-three cases of single text visualization and twenty-six cases of collections of texts, out of a total number of forty-nine reviewed cases. See Figure ??.

Figure 9. Reviewed cases in numbers



(a)



(b)

Report on question 3: Does the visualization use elements from the discourse structure or from the syntactic structure? Eight cases are classified as not-applicable because those are neither following the discourse structure, nor the syntactic one. In general, the cases represent keywords from the text independently of their function within the text.

- Report on question 8: Is it valid for small or large datasets? The reason for the ten N/A results is that those cases are independent from the size of the dataset. Three of the cases that Jeff Clark did using the text “Les Miserables” by Victor Hugo (cases 1, 16, and 24) and the case “Phrase Nets” (case 19) by van Ham, Wattenberg and Viegas are a product of text analysis, and the size of the text is not directly related to the actual

visualizations. The “State of the Union 2011 - Sentence Bar Diagrams” also by Jeff Clark (case 4) is a sentence coloured approach that, independently of the length of the text, shows the main topic of each sentence. It is a similar reason for the cases of “Poem Viewer” (case 3), which explores and analyses poems word by word, and “Literature fingerprint” by Keim and Öлке, which allows comparison of writing styles based on sentence lengths. The two cases “Visualizing Lexical Novelty in Literature” by Matthew Hurst (case 5) and “On the Origin of Species: The Preservation of Flavored Traces” by Ben Fry (case 6) are not dependent directly on the dataset size for other reasons. In those cases, the evolution of a text, following the sequence of the text, allows long text representation. Finally, the case “Pediometer” by Mueller-Birn, Benedix and Hantke (case 42) shows Wikipedia edits on time through a physical device, and this is also not related to the size of the Wikipedia.

Table 2. Number of classification questions, and not applicable (N/A) answers.

Group	Question ID	No of answers	No of N/A	% of N/A
Single	1	23	-	-
Single	2	23	-	-
Single	3	23	8	34.8
Collection	4	26	-	-
Collection	5	26	-	-
Both	6	49	-	-
Both	7	49	-	-
Both	8	49	10	20.4
TOTAL		268	18	6.7

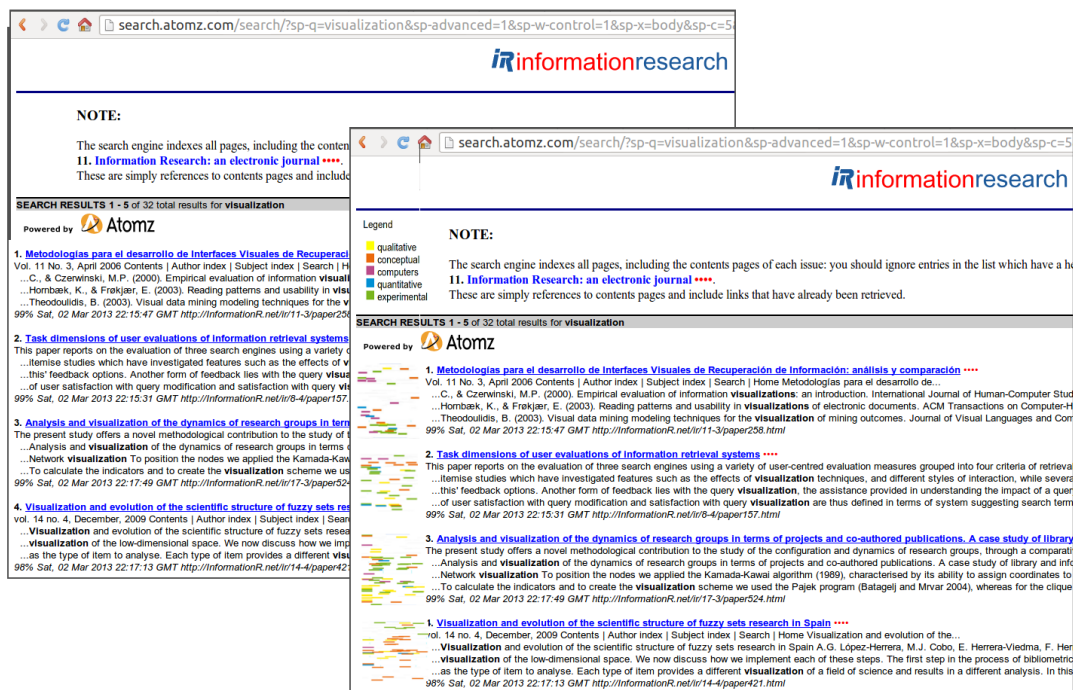
5.2. Search results and single text visualization

According to the classification schema presented, the visualization tool proposed to aid selection of texts from search results outputs, called Texty, is classified as: whole text visualization. It follows the sequence of the text, and neither follows a discourse or syntactic structure. The datasets are static, not search related. It is not suitable for big datasets either.

I generated Textys for every paper in the Information Research journal, from Volume 1, No. 1 (1995) to Volume 15 No.4 (2010), with a total of 454 Textys.

A Texty demo for the Information research papers is accessible online. The access is restricted (To access, use user='texty' and password='texty') (Jaume Nualart, 2012, a). The five vocabularies defined and the terms for each vocabulary are also accessible (Jaume Nualart, 2013).

Figure 10. Example of textys incorporated to a traditional flat list of search engine results. In the top-left there is the legend of the five textys colors, which is required to read the textys attached to each search result item.




















Analysis of an issue of the journal

Figure ?? shows seventeen Textys representing an issue of the journal Information Research. This representation can help the reader to select papers as follows:

- The predominant vocabulary in this issue is experimental (green), though followed closely by qualitative approach (yellow).
- Three of the represented papers (numbers 3, 11 and 13) look clearly experimental (green), while paper 7 looks like one that requires more computers/IT's readers knowledge (violet).
- Five of the seventeen papers (38.5%) have a notable presence of computer/IT (violet).
- The paper with the biggest conceptual load can be identified visually (number 9), despite that there are three others (number 7, 8 and 16) that also have a conceptual content (orange).

- The more generalist paper seems to be the number 15. This issue does not involve quantitative methodologies much (blue).

Figure 11. Textys for Section volume 15, No 4 (December 2010) of the Information Research and legend.

<p>1. <i>Proceedings of ISIC</i></p>  <p><i>Cultural differences in the health information environments and practices between Finnish and Japanese university students</i> Graeme Baxter, Rita Marcella and Laura Illingworth</p>	<p>2. <i>Proceedings of ISIC</i></p>  <p><i>Organizational information behaviour in the public consultation process in Scotland</i> Leanne Bowler</p>	<p>3. <i>Proceedings of ISIC</i></p>  <p><i>Talk as a metacognitive strategy during the information search process of adolescents</i> Jenny Bronstein</p>	<p>4. <i>Proceedings of ISIC</i></p>  <p><i>Selecting and using information sources: source preferences and information pathways of Israeli library and information science students of your paper</i> Donald O. Case</p>
<p>5. <i>Proceedings of ISIC</i></p>  <p><i>A model of the information seeking and decision making of online coin buyers</i> Kreetta Askola, Toshimori Atsushi and Majje-Leena Huotari</p>	<p>6. <i>Proceedings of ISIC</i></p>  <p><i>Local versus global information relevance in Website use: a case study with the information literacy portal ALINKES</i> Francisco Javier García Marco and Maria Pinto</p>	<p>7. <i>Proceedings of ISIC</i></p>  <p><i>Information behaviour research and information systems development: the SHAMAN project, an example of collaboration</i> Elena Maceviciute and T.D. Wilson</p>	<p>8. <i>Proceedings of ISIC</i></p>  <p><i>Avoiding health information in the context of uncertainty management</i> Anu Sairanen and Reijo Savolainen</p>
<p>9. <i>Proceedings of ISIC</i></p>  <p><i>A study of labour market information needs through employers' seeking behaviour</i> Sonia Sanchez-Cuadrado, Jorge Morato and Yorgos Andreidakis</p>	<p>10. <i>Proceedings of ISIC</i></p>  <p><i>Information in context: co-designing workplace structures and systems for organizational learning</i> Mary M. Sonerville and Zaena Howard</p>	<p>11. <i>Proceedings of ISIC</i></p>  <p><i>"We have a lot of information to share with each other": Understanding the value of peer-based health information exchange</i> Tiffany C. Veinot</p>	<p>12. <i>Proceedings of ISIC</i></p>  <p><i>Information sharing: an exploration of the literature and some propositions</i> T.D. Wilson</p>
<p>13. <i>Proceedings of ISIC</i></p>  <p><i>Applying McKenzie's model of information practices in everyday life information seeking in the context of the menopause transition</i> Alison Yeoman</p>	<p>14. <i>Regular paper</i></p>  <p><i>Double or nothing: is redundancy of spatial data a burden or a need in the public sector of Uganda?</i> Walter T de Vries and Beatrice Winnie Nyirama</p>	<p>15. <i>Regular paper</i></p>  <p><i>Analysis of automatic translation of questions for question answering systems</i> Loia García-Santiago and María-Dolores Olvera-Lobo</p>	<p>16. <i>Regular paper</i></p>  <p><i>Dietary blogs as sites of informational and emotional support</i> Reijo Savolainen</p>
<p>17</p>  <p><i>Information and information science: an address on the occasion of receiving the award of Doctor Honoris Causa, at the University of Murcia, 30 September, 2010</i> T.D. Wilson</p>			<p>Legend</p> <ul style="list-style-type: none"> ■ qualitative ■ conceptual ■ computers ■ quantitative ■ experimental

Comparative with classic systems: bar chart, line chart

The objective of a comparison of Texty and classic systems is not to quantitatively study Texty's performance against other alternative visualizations. The objective is to show the extra features that Texty offers in relation to the charts:

- Texty shows the distribution of terms along the text, which is in each part of the physical text.
- With Texty, the conceptual structure of the paper can be seen, e.g.: at the start of the paper there is a conceptual explanation; then the experimental part is developed; finally, the calculations in which there is intensive use of technology related and/or computer related operations.
- Texty does not need axes or coordinates and scales.

5.3. Digital libraries and text collections visualization

The application created to explore and visualize Information Research published papers, called Area, is classified as text collection visualization. The items of the collection are visually differentiated by coloured squares. The squares represent static datasets. The items can be searched and filtered according to configuration of each field. Finally, the number of items that can be represented is not considered big data.

The code of this application can be found online [Jaume Nualart, 2015a, b].

Digital library visualization results

The ten features considered for implementation in the new interface are listed in three groups:

- a) Existing features: no implementation done, but redirection to the existing feature of Information Research existing interface.
 - Explore by issue as a list of papers.
 - Search with Atomz, and search with Google.
- b) Improved features: features that exist in both interfaces, but the features of Area bring some improvements.
 - Explore by Year, Issue and Volume.

- Explore by Subject.
 - Explore by Author (authors can have more than one paper).
 - How many papers talk about a subject?
- c) New features: features not present in the existing interface and implemented as new by the Area interface.
- Multiple overviews of the collection
 - Numerical overview
 - Topic distribution
 - Explore by language

For each item of the collection represented, i.e., for each paper, thirteen fields are included as metadata. See Table 1 from Paper III.

User evaluation results

I asked participants to compare seven tasks completed with the journal's existing interface on the one hand, and with Area's interface on the other, and state which of the two tools suited them better in their opinion. The survey was answered by forty-four participants. Around 80% of participants have chosen Area interface, while about 10% chose the existing interface, and 10% found no significant difference. The results for each task proposed in the survey are commented below:

- How many papers have been published in the journal since the first issue?

The lack of overview numbers (e.g., total number of papers in the collection) in digital collections is usual. There is no need of visualization techniques for this feature, and its implementation is trivial. I added this minimal information that gives an idea of the size of the represented collection: Number of papers in IR = 592. Total of 74 issues in twenty years (1995-2015).

83.78% of the participants preferred Area interface for this task.

- How many papers talk about visualization?

The DL that offer search and filter items do not always state the weight of the results in relation to the whole collection. In my proposal, the user receives the number of results for each query, the proportion that the number of results represents in the entire collection, and its visual distribution along the chosen parameters.

(64.86% prefer Area and a 27.03% finds no difference)

- Understanding the topics and themes of the journal.
Area shows the distribution of the subjects defined by Information Research editors along the collection.
(83.78% prefer Area, and a 10.81% prefer the existing interface)
- Exploring new topics and discovering new research in this field.
The feature of multiple grouping of Area could make the process of discovering papers of the collection easier.
(78.38% prefer Area, and 13.51% prefer the existing interface)
- How is the term “visualization” distributed in the history of the journal?
Area shows the distribution of search results along the collection.
(83.78% prefer Area interface, and 13.51% did not find any difference between both interfaces)
- When exploring papers of the journal website: do you have a better overview of the journal using the existing interface or the Area interface?
This is a main feature of the Area interface.
(81.08% prefer Area, and 16.22% prefer the existing interface)
- Finding papers related to your personal interests.
In this subjective task, the participants also show preference for the proposed interface.
(81.08% prefer Area interface, 10.81% find no difference, and 8.11% prefer the existing interface)

Performance evaluation results

The implementation of this application is constrained by a number of items in the dataset and the dataset size. The first of these is related to screen resolution, while the second is related to the size of RAM memory available on the client side. Performance tests conducted suggest the use of collections that do not exceed fifty thousand items. The performance tests are accessible online (Jaume Nualart, 2015b).

6. Discussion and conclusions

This research aims to improve the way humans work with textual documents when doing tasks such as exploring, discovering, searching, filtering, collecting, indexing, comparing, or just reading. From theory to practice, I present a thesis as a compendium of three complementary publications that incorporate three tools to the scientific literature: a classification schema for text visualization approaches, an approach that represents individual textual documents, and a visual interface for digital libraries. Below I present a list of conclusions with attached short discussions.

The elaboration of a classification schema has been based on the visual features and reviewed approaches instead of their task solving goals. The reason for this is that a defined task does not imply a visual method to help to accomplish that task. Very different visualization strategies can apply to a single task. Therefore, since I am classifying text visualization ideas, I decided to compare their visual baseline strategies when representing a dataset, instead of their specific uses and real scenario applications.

I have answered the question “How to classify text visualization cases according to visual features?” presenting a classification schema as a research tool. It can be used as a whole or in parts. Any of the questions in the schema can be used individually, that is: text sequence, discourse or syntactic structures, items or aggregates, data landscape, big or small dataset, static or dynamic dataset, etc. Once those binary options are assimilated by a community, they can become a standard. Standardisation brings consistency, a universality in concepts, and I consider this fact remarkable when I take into account that the field of text visualization is young.

When writing this thesis, Paper I was cited by another text visualization review paper (Kucher and Kerren, 2015). In this paper, the Kucher and Kerren study collected one hundred and forty cases that can be explored with an online interface. The cases were collected with contributions of visitors in a crowdsourced style. This review includes only cases published in academic journals. Applying my methodology, I plan to classify those cases in order to continue testing the consistency of the proposed classification schema. It is also possible to compare both schemes. They use a very wide classification schema. Eight classification axes are defined: analytic tasks, visualization tasks, data source, data properties, data domain, visualization dimensionality, visualization representation and visualization alignment. Each category includes non-restrictive subcategories, so each case can have multiple subcategories of the same category. The Kucher and Kerren classification is very inclusive and probably more exposed to subjectivity than the schema I propose, which is much more exclusive and univocal tendentious.

From the application of the classification schema, to the reviewed cases, I can state some conclusions:

- The most popular and assimilated text visualizations techniques are word clouds and network diagrams for part text visualization, and treemaps for collections of aggregations.
- Single-text visualizations have been applied mainly to literature, a field that, apart from being characterised by complex combinations of words, can present high levels of human abstraction and freedom of structure and experimentation.
- I have identified only one single/partial-text approach that is sequential (Document arc diagrams, case 22). Most partial-text visualizations extract the essence of the text based on one or more criteria and so the original sequence of the text is lost. Since sequential visualization approaches present certain advantages, it seems that partial-visualization approaches that maintain the original text sequence should be encouraged.
- Text-collection visualizations tend to employ methods that are used for data visualization in general.
- Text collection aggregations is the category in which the most specific designs and ideas have been developed. More work needs to be undertaken to identify any common approaches in this kind of visualization.

In the future, I want to conduct an evaluation of the proposed classification with a survey to compare several classifications schemes.

The second challenge I presented in this dissertation is represented by the idea of Texty in Paper II and the question “how can we improve the traditional flat list of search engine results using visualization techniques and not interfering with information retrieval operations?”. The answer is Texty, a single text visualization idea that can be applied not only to search results – as shown in Figure 10–, but also to any flat list of documents. In Figure 11, I show an example of this: an issue of the Information Research journal with seventeen papers and their Textys. As explained in Chapter 5 the journal issue with Textys brings more accurate information about its contents, giving an overview of the issue as well as individual view of each published paper in the issue, and it may help to take decisions about which papers to select for reading first and which ones not to select at all.

Texty is a tool that, from a technological point of view, can be implemented in an existing collection of texts in a non-intrusive way. That means that, in most cases, it will not be necessary to change or reprogram the storage and information system where the collection of texts are held.

Future planned developments include a user evaluation study with Textys integrated in an in-production search engine or DL.

For the third challenge, the study of visual DL interfaces, I tried to accomplish most the conditions defined in Chapter 3 Aims and research questions (“how to explore and search in DL using an interface based on graphic elements and visual language instead of only-textual interfaces?”): a compromise between the learning curve for new users when facing a new interface, and the step forward to innovation in DL interfaces. The architectural proposal articulates, in a single structure, the two main systems facilitating the location of information in digital contexts, i.e., the navigation system and the search system. These two systems are present in most contexts, which is a guarantee that users are fully familiar with them and that additional specific instructions are not needed for them to use Area efficiently and comfortably.

The proposed idea, materialised as a software called Area, also provides feasibility of the implementation of in-production DL, allowing remote and local installations in a non-intrusive way. This ease in the technical aspects also increases feasibility in terms of budget and time.

The user test conducted on Area provided a positive response: users detect and understand the new interface and its new features in relation to the existing interface. Users prefer or require new ways of presenting information, and users feel confident and positive about using the new features.

Future research includes new applications of the tools to other data collections, and, thus, a continuous development of the tool, improving performance and usability. Also, it could be interesting to build an interface combining the two artifacts in one interface, Texty and Area.

Bibliography

- Terry J Anderson, Ali Hussam, Bill Plummer, and Nathan Jacobs. Pie charts for visualizing query term frequency in search results. In *Proceedings of the 5th International Conference on Asian Digital Libraries: Digital Libraries: People, Knowledge, and Technology*, ICADL '02, pages 440–451, London, UK, UK, 2002. Springer-Verlag. ISBN 3-540-00261-8. URL <http://dl.acm.org/citation.cfm?id=646228.681545>.
- Keith Andrews, Christian Gutl, Josef Moser, Vedran Sabol, and Wilfried Lackner. Search result visualisation with xfind. In *User Interfaces to Data Intensive Systems, 2001. UIDIS 2001. Proceedings. Second International Workshop on*, pages 50–58. IEEE, 2001.
- Gary J Anglin, Hossein Vaez, and Kathryn L Cunningham. Visual representations and learning: The role of static and animated graphics. *Handbook of research on educational communications and technology*, 2:865–916, 2004.
- Ricardo Baeza-Yates. Tendencias en recuperación de información en la web. *Bid*, (27):1–4, 2011.
- Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.
- Ricardo Baeza-Yates, Andrei Z Broder, and Yoelle Maarek. The new frontier of web search technology: seven challenges. In *Search computing*, pages 3–9. Springer, 2011.
- Marcia J Bates. The design of browsing and berrypicking techniques for the online search interface. *Online review*, 13(5):407–424, 1989.
- Sabine Bauer. Interactive visualizations for search processes. 2014.
- David Benavides, Sergio Segura, and Antonio Ruiz-Cortés. Automated analysis of feature models 20 years later: A literature review. *Information Systems*, 35(6):615–636, 2010.
- Matthew Chalmers and Paul Chitson. Bead: Explorations in information visualization. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 330–337. ACM, 1992.
- Christopher Collins, Sheelagh Carpendale, and Gerald Penn. Docuburst: Visualizing document content using language structure. In *Computer Graphics Forum*, volume 28, pages 1039–1046. Wiley Online Library, 2009.

- John V Cugini, Sharon Laskowski, and Marc M Sebrechts. Design of 3d visualization of search results: evolution and evaluation. In *Electronic Imaging*, pages 198–210. International Society for Optics and Photonics, 2000.
- Mary Czerwinski, Susan Dumais, George Robertson, Susan Dziadosz, Scott Tiernan, and Maarten Van Dantzich. Visualizing implicit queries for information management and retrieval. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 560–567. ACM, 1999.
- Susan Dziadosz and Raman Chandrasekar. Do thumbnail previews help users make better relevance decisions about web search results? In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 365–366. ACM, 2002.
- Education Ltd TLS. World university rankings 2014-2015 - times higher education. <https://www.timeshighereducation.co.uk/world-university-rankings/2015/world-ranking>, 2015. URL <https://www.timeshighereducation.co.uk/world-university-rankings/2015/world-ranking>.
- D E Egan, J R Remde, L M Gomez, T K Landauer, J Eberhardt, and C C Lochbaum. Formative design evaluation of superbook. *ACM Transactions on Information Systems (TOIS)*, 7(1):30–57, 1989.
- Edward A Fox, Deborah Hix, Lucy T Nowell, Dennis J Brueni, William C Wake, Lenwood S Heath, and Durgesh Rao. Users, user interfaces, and objects: Envision, a digital library. *Journal of the American Society for Information Science*, 44(8):480–491, 1993.
- M. Grobelnik and D. Mladenic. Efficient visualization of large text corpora. In *Proceedings@articlebauerinteractive, title=Interactive Visualizations for Search Processes, author=Bauer, Sabine dings of the seventh seminar. Dubrovnik, Croatia, 2002*. URL <http://ailab.ijs.si/dunja/SiKDD2002/papers/GrobelnikSep02.pdf>.
- Marti Hearst. Search user interfaces. *Search User Interfaces*, 54(Ch 1):404, November 2009. ISSN 00010782. doi: 10.1145/2018396.2018414. URL <http://searchuserinterfaces.com/book/>.
- Marti A Hearst and Chandu Karadi. Cat-a-cone: an interactive interface for specifying searches and viewing retrieval results using a large category hierarchy. In *ACM SIGIR Forum*, volume 31, pages 246–255. ACM, 1997.
- Jeff Heer. A conversation with jeff heer, martin wattenberg, and fernanda viegas, 2010.
- Daniel Hienert, Frank Sawitzki, Philipp Schaer, and Philipp Mayr. Integrating interactive visualizations in the search process of digital libraries and ir systems. In *Advances in Information Retrieval*, pages 447–450. Springer, 2012.

- O Hoeber and X D Yang. A comparative user study of web search interfaces: HotMap, concept highlighter, and google. In *Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on*, pages 866–874, 2006.
- K Hornbaek and E Frokjaer. Reading of electronic documents: the usability of linear, fisheye, and overview+ detail interfaces. In *Conference on Human Factors in Computing Systems: Proceedings of the SIGCHI conference on Human factors in computing systems*, volume 2001, pages 293–300, 2001.
- Information Research journal. Information research, 2015. URL <http://www.informationr.net/ir/>. Accessed: 08/08/2015.
- Jaume Nualart. Texty representations of Journal Information Research papers, 2012. URL <http://o.subvideo.tv/>.
- Jaume Nualart. Texty terms, 2013. URL <http://vis.subvideo.tv/texty/terms.php>.
- Jaume Nualart. Area code, 2015a. URL <https://github.com/jaumet/Area>. [Accessed: 2015-01-03].
- Jaume Nualart. Area stress test, 2015b. URL <http://research.nualart.cat/area-stress/>. [Accessed: 2015-01-03].
- Jaume Nualart. Area representations text visualization approaches, 2015c. URL <http://research.nualart.cat/area-ir/>.
- N Jhaveri and K J Raeihae. The advantages of a cross-session web workspace. In *CHI'05 extended abstracts on human factors in computing systems*, pages 1949–1952, 2005.
- Shaun Kaasten, Saul Greenberg, and Christopher Edwards. How people recognise previously seen web pages from titles, urls and thumbnails. In *People and Computers XVI-Memorable Yet Invisible*, pages 247–265. Springer, 2002.
- SA KartOO. Kartoo. 2009. URL Online:<http://www.kartoo.de> [Stand:2005-07-10].
- Beomjin Kim, Jon Scott, and Seung Eun Kim. Exploring digital libraries through visual interfaces. 2011.
- Kostiantyn Kucher and Andreas Kerren. Text visualization techniques: Taxonomy, visual survey, and community insights. In *8th IEEE Pacific Visualization Symposium (PacificVis' 15), Hangzhou, China*, pages 117–121. IEEE Computer Society, 2015.
- Frederick Wilfrid Lancaster. Vocabulary control for information retrieval. 1972.
- W Howard Levie and Richard Lentz. Effects of text illustrations: A review of research. *ECTJ*, 30(4):195–232, 1982.

- Thomas M Mann. Visualization of www-search results. In *Database and Expert Systems Applications, 1999. Proceedings. Tenth International Workshop on*, pages 264–268. IEEE, 1999.
- Noah Iliinsky. *Choosing visual properties for successful visualizations*. s IBM Software - Business Analytics, 2013. URL <http://public.dhe.ibm.com/common/ssi/ecm/en/ytw03323usen/YTW03323USEN.PDF>.
- Jaume Nualart-Vilaplana, Mario Pérez-Montoro, and Mitchell Whitelaw. How we draw texts: A review of approaches to text visualization and exploration. *El profesional de la información*, 23(3):221–235, 2014.
- Mario Pérez-Montoro. Arquitectura de la información en entornos web. *El profesional de la información*, 19(4):333–338, 2010.
- George G Robertson, Jock D Mackinlay, and Stuart K Card. Cone trees: animated 3d visualizations of hierarchical information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 189–194. ACM, 1991.
- Alexander Shaphiro. Meatball wiki: Touchgraph, 2002. URL <http://meatballwiki.org/wiki/TouchGraph>. Accessed: 08/08/2015.
- Alexander Shapiro. Touchgraph, 2003.
- B Shneiderman. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on graphics (TOG)*, 11(1):92–99, 1992.
- Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pages 336–343, 1996.
- Ben Shneiderman, David Feldman, Anne Rose, and Xavier Ferré Grau. Visualizing digital library search results with categorical and hierarchical axes. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 57–66. ACM, 2000.
- Jacqueline Strecker and Tricia Wind. *Data visualization in review: summary*. IDRC, Ottawa, ON, 2012.
- Wilko Van Hoek and Philipp Mayr. Assessing visualization techniques for the search process in digital libraries. *arXiv preprint arXiv:1304.4119*, 2013.
- Wilko van Hoek and Philipp Mayr. Is evaluating visual search interfaces in digital libraries still an issue? *arXiv preprint arXiv:1408.5001*, 2014.
- T.D. Wilson. Information research editorial vol 18 no 2, 2013. URL <http://www.informationr.net/ir/18-2/editor182.html>. Accessed: 08/08/2015.

Allison Woodruff, Andrew Faulring, Ruth Rosenholtz, Julie Morrision, and Peter Pirollo. Using thumbnails to search the web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '01, pages 198–205, New York, NY, USA, 2001. ACM. ISBN 1-58113-327-8. doi: 10.1145/365024.365098. URL <http://doi.acm.org/10.1145/365024.365098>.

A. Appendix: Published papers

A.1. Paper I

Citation:

Nualart-Vilaplana, Jaume; Pérez-Montoro, Mario; Whitelaw, Mitchell (2014). How we draw texts: a review of approaches to text visualization and exploration. *El profesional de la información*, mayo-junio, v. 23, n. 3, pp. 221-235.

Afiliation (in catalan)

- 1r autor: Jaume Nualart Vilaplana es doctorand a la Facultat de Biblioteconomia i Documentació de la Universitat de Barcelona, inginyer de recerca al NICTA (Australia), i doctorand a la Faculty of Arts and Design, University of Canberra (Australia).
- 2n autor: Mario Pérez-Montoro, doctor i professor al Dept. de Ciències de la Informació de la Universitat de Barcelona.
- 3r autor: Mitchell Whitelaw, doctor i professor associat de la Faculty of Arts and Design de la University of Canberra

Summary (in Catalan)

En aquest treball es presenta una revisió d'estratègies per a la visualització i exploració de textos, argumentant que la visualització de textos constitueix un subcamp de la visualització de dades que es nodreix dels avanços en l'anàlisi de textos i de la creixent quantitat de dades accessibles en format text. Proposem una classificació original per a un total de quaranta-nou casos revisats. La classificació està basada en les característiques visuals de cada cas, identificades mitjançant un procés inductiu d'anàlisi. Agrupem els casos (publicats entre el 1994 i el 2013) En dues categories: visualització de textos individuals i visualització de col·leccions de textos. Els casos revisats poden ser explorats i comparats en línia.

Access online (URL):

- <http://dx.doi.org/10.3145/epi.2014.may.02>

- <http://www.elprofesionaldelainformacion.com/contenidos/2014/may/02.pdf>

ARTÍCULOS



HOW WE DRAW TEXTS: A REVIEW OF APPROACHES TO TEXT VISUALIZATION AND EXPLORATION



Jaume Nualart-Vilaplana, Mario Pérez-Montoro y Mitchell Whitelaw

Nota: Este artículo puede leerse traducido al español en:

http://www.elprofesionaldelainformacion.com/contenidos/2014/may/02_esp.pdf



Jaume Nualart-Vilaplana is a PhD candidate in the *Faculty of Arts and Design, University of Canberra* (Australia), research engineer at *Nicta* (Australia), and a PhD candidate in the *Faculty of Information Science, University of Barcelona*. MAS and MSc (Licenciatura) at *Autonomous University of Barcelona*

<http://orcid.org/0000-0003-4954-5303>

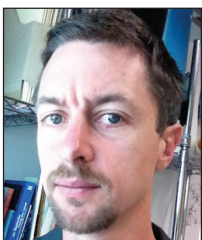
*Machine Learning Research Group at NICTA, Canberra Research Laboratory
Tower A, 7 London Circuit, Canberra City ACT 2601, Canberra, Australia
jaume.nualart@canberra.edu.au*



Mario Pérez-Montoro holds a PhD in Philosophy and Education from the *University of Barcelona* and a Master in Information Management and Systems from the *Polytechnic University of Catalonia*. He studied at the *Istituto di Discipline della Comunicazione* at the *Università di Bologna* (Italy) and has been a visiting scholar at the *Center for the Study of Language and Information (CSLI)* at *Stanford University* (California, USA) and at the *School of Information at UC Berkeley* (California, USA). He is a professor in the *Department of Information Science* at the *University of Barcelona*. His work has focused on information architecture and visualization. He is author of the book *Arquitectura de la información en entornos web* (Trea, 2010).

<http://orcid.org/0000-0003-2426-8119>

*Facultat de Biblioteconomia i Documentació, Universitat de Barcelona
Melcior de Palau, 140. 08014 Barcelona, España
perez-montoro@ub.edu*



Mitchell Whitelaw is an academic, writer and practitioner with interests in new media art and culture, especially generative systems and data-aesthetics. His work has appeared in journals including *Leonardo*, *Digital creativity*, *Fibreculture*, and *Senses and society*. In 2004 his work on a-life art was published in the book *Metacreation: art and artificial life* (MIT Press, 2004). His current work spans generative art and design, digital materiality, and data visualisation. He is currently an associate professor in the *Faculty of Arts and Design* at the *University of Canberra*, where he leads the *Master of Digital Design*. He blogs at *The Teeming Void*.

<http://orcid.org/0000-0001-9013-9732>

*Faculty of Arts and Design, University of Canberra
Bldg, Floor & Room: 9, C12. ACT 2617, Canberra, Australia
mitchell.whitelaw@canberra.edu.au*

Abstract

This paper presents a review of approaches to text visualization and exploration. Text visualization and exploration, we argue, constitute a subfield of data visualization, and are fuelled by the advances being made in text analysis research and by the growing amount of accessible data in text format. We propose an original classification for a total of 49 cases based on the visual features of the approaches adopted, identified using an inductive process of analysis. We group the cases (published between 1994 and 2013) in two categories: single-text visualizations and text-collection visualizations, both of which can be explored and compared online.

Keywords

Review, Text visualization, Data visualization, Data exploration, Data display, Information visualization, Text analysis.

Título: Cómo dibujamos textos. Revisión de propuestas de visualización y exploración textual

Article received on 19-01-2014

Approved on 09-03-2014

Resumen

En este trabajo se presenta una revisión de estrategias para la visualización y exploración de textos. Se argumenta que la visualización y exploración de textos constituye un subcampo de la visualización de datos que se nutre de los avances en el análisis de textos y de la creciente cantidad de datos accesibles en formato texto. Proponemos una clasificación original para un total de cuarenta y nueve casos revisados. La clasificación está basada en las características visuales de cada caso, identificadas mediante un proceso inductivo de análisis. Agrupamos los casos (publicados entre 1994 y 2013) en dos categorías: las visualizaciones de texto individuales y la visualizaciones de colecciones de textos. Los casos revisados pueden ser explorados y comparados en línea.

Palabras clave

Visualización de texto, Visualización de datos, Exploración de datos, Visualización de información, Análisis de textos.

Nualart-Vilaplana, Jaume; Pérez-Montoro, Mario; Whitelaw, Mitchell (2014). "How we draw texts: a review of approaches to text visualization and exploration". *El profesional de la información*, mayo-junio, v. 23, n. 3, pp. 221-235.

<http://dx.doi.org/10.3145/epi.2014.may.02>

1. Introduction

The aim of this review is to propose a classification of text visualization and exploration tools, while describing the broader context in which they operate. To do so, we list, classify and discuss the most important contributions made in the field of text visualization and exploration between 1994 and 2013. This field is undergoing rapid growth –fuelled by open data initiatives and web scraping– and has become highly diversified, developing in parallel in a range of disciplines. Some of the most important visualization methods invented between 1765 and 1999 were the timeline, bar chart, pie chart, flow map, Venn diagram, histogram, Gantt chart, flowchart, tag cloud, social networks, boxplot, star plot, treemap, headmap, and sparkline. Figure 1 presents a word cloud (using *Wordle*) of the professions practiced by their respective inventors. Given this diversity, our search for cases has been conducted in many different contexts and has involved the examination of many different sources, ranging from the sciences to the humanities, from academic journals to blog sites, from universities to freelance studios, and from open data institutions to open data communities. Clearly this proliferation of disciplines has meant the adoption of a variety of different philosophies and points of view.

This review aims to help those that work with data, and especially with texts (but by no means limited to academics), to use visualization techniques that can identify patterns or behaviours present in the textual reality. Moreover, these techniques can help users improve –in terms of both the

speed and the clarity of the process– the way in which they visualize and discover the facts that lie within the data.

Drawing a clear conceptual line between approaches to text visualization and exploration is no straightforward task, but here we have opted to review cases dedicated to both processes, be they described separately or together. Note that on occasions, for the sake of simplicity, we use the term text visualization in reference to both approaches.

The two types of text visualization considered here are:

1) Single-text representation, that is, ways of extracting meaning from texts based on writing style, document structure and language register as opposed to pure statistics. Our interest lies in representing the meaning and salient features of texts because their convenient visualization can speed up and/or improve our ability to select texts and manage the time required to tackle them. The research output of fields such as natural language processing, linguistic computing and machine learning provides techniques for producing high quality data representing complex texts. It is our belief that by combining these techniques with a suitable text visualization method we can improve the way in which we examine and understand texts.

2) Representation and exploration of collections of texts. Exploring and selecting individual texts and navigating and analyzing collections of texts are daily tasks for many of those who work with computers and datasets, and there is clearly plenty of room for new ideas and tools to facilitate

their work. Information retrieval is a critical factor in an environment characterized by an excess of information (Baeza-Yates *et al.*, 1999). When a user conducts a search, the information retrieval systems normally respond with a list of results. More often than not, the presentation of these results plays an important role in satisfying the user's information needs, so a poor or inad-

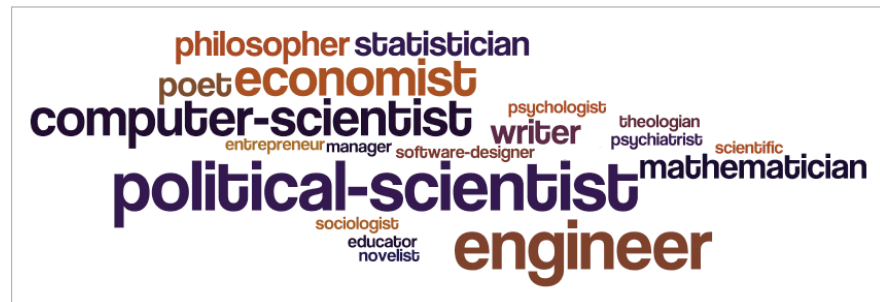


Figure 1. Word cloud of the professions practiced by inventors of visualization methods

Table 1. Leading universities and their data visualization departments

Institution	Rank in 2012	Department/Course	URL
Harvard University	1	Broad Institute of Harvard and MIT	http://www.broadinstitute.org/vis
Massachusetts Institute of Technology	2	Broad Institute of Harvard and MIT	http://www.broadinstitute.org/vis
University of Cambridge	3	--	--
Stanford University	4	Stanford Vis Group	http://vis.stanford.edu
University of California, Berkeley	5	VisualizationLab	http://vis.berkeley.edu
University of Oxford	6	Visual Informatics Lab at Oxford	http://oxvi.wordpress.com
Princeton University	7	PrincetonVisLab	http://www.princeton.edu/researchcomputing/vis-lab
University of Tokyo	8	--	--
University of California, Los Angeles	9	IDRE GIS and visualization	https://idre.ucla.edu/visualization
Yale University	10	--	--

Table 2. Conferences dedicated primarily to data visualization ordered by number of participants (Stefaner, 2013)

Conference	Location	Topic	No. participants	URL
Nicar 2013	USA	Data journalism	149	http://ire.org/conferences/nicar-2013
Dd4d 2009	France	Information visualization	52	http://www.dd4d.net
FutureEverything 2013	UK	Technology/society/art	52	http://futureeverything.org
Resonate 2013	UK	Creative code	44	http://www.thisisresonate.co.uk/resonate-13
Graphical web 2012	Switzerland	Open web/datavis	38	http://www.graphicalweb.org/2012
IeeeVis - VisWeek 2012	USA	Information visualization	-	http://ieevis.org
EuroVis 2013	Germany	Computational aesthetics	-	http://www.eurovis2013.de
Siggraph 2013	USA	Computer graphics and interactive techniques	-	http://s2013.siggraph.org
OzViz 2012	Australia & NZ	Workshops for visualisation practitioners, academics and researchers	-	http://www.ozviz2012.org

equate presentation can thwart the user (Baeza-Yates *et al.*, 2011). Typically, information retrieval systems present the results of a query in a flat, one-dimensional list. Such lists tend to be opaque in terms of the order they give to the information, i.e., the users are unaware as to why the list is presented in a particular order. To refine their search, users have to interact again, normally by filtering the first output of results. It is our belief that new techniques for representing collections of texts –including search results– can help improve navigation, exploration and retrieval.

As we show below, data visualization can today be considered a consolidated academic field (Strecker; IDRC, 2012). Thus:

- Seven of the top 10 universities according to the *Times Higher Education ranking* (2012) have departments or research groups working in the field of data visualization. The discipline is incorporated in a wide variety of departments, ranging from computer science and statistics to linguistics and graphical design, and from chemistry and physics to genetics and history. Recently, data visualization has emerged as a distinct field, with specific departments dedicated to its study and master's programs being taught in the subject (table 1).

- Over the last five years a number of conferences have been dedicated primarily to data visualization. These are listed in table 2.

- A number of journals are now specifically dedicated to studies in data visualization, and important contributions can be found also in conference proceedings (table 3).

Finally, a number of leading websites –including *Infosthetics*, *Visualcomplexity* and *Visualizingdata.com*– play a key role in the dissemination of the subject.

1.1. Text visualization

Shneiderman (1996) classifies regular texts as one-dimensional data, that is, data organized in a sequential manner, running right-to-left (or left-to-right), line-by-line, top-to-bottom. Yet, a text can have multiple internal structures, a morphology made up of paragraphs, sentences and words.

Table 3. Main journals dedicated to data visualization

Name	Url
Parsons journal for information mapping	http://pjim.newschool.edu/issues/index.php
Journal of visualization	http://springer.com/materials/mechanics/journal/12650
Ieee Transactions on visualization and computer graphics (TVCG)	http://www.computer.org/portal/web/tvcg
Information visualization	http://ivi.sagepub.com
International journal of image processing and data visualization (Ijipdv)	http://iartc.net/index.php/Visualization
IEEE Vis (former Visweek)	http://ieevis.org
EuroVis	http://www.eurovis2013.de
ACM CHI	http://chi2013.acm.org
EG CGF	http://www.eg.org
IVS	http://www.graphiclink.co.uk/IV2013

Depending on its information structure, a text may be ordered by chapters, parts, sections, subsections, etc. If a text is given in a specific format, such as html, then it may be organized into bodies, divs, paragraphs, etc. In these examples the text includes tree structures as well as a one-dimensional structure. Additionally, texts may have a subjective component and an abstract structure that is not readily analysed by a computer. All in all, these data types and structures constitute the specificities of a text.

The amount of data to which we have access grows on a daily basis. Most of these data are in text format, as **Fernanda Viégas** and **Martin Wattenberg** in an interview with **Jeff Heer** argue: “One of the things I think is really promising is visualizing text. That has been mostly ignored so far in terms of information visualization approaches, and yet a lot of the richest information we have is in text format” (**Heer**, 2010).

Seven of the top 10 universities have departments or research groups working in data visualization

Data analysis defines the boundaries of data visualization, i.e., it provides the fine line between multiple truths and lies. In the case of text visualization, this role has been taken on by text analysis: in the main, via computational linguistics, natural language processing, machine learning and statistics. The advances made in text analysis at a whole range of levels have provided computers with text understanding, enabling them to modify a text, the so-called unstructured data (see next subsection “Text analysis”).

There is some discussion as to whether text visualization might be considered a specific subfield of data visualization. Some authors tend to disagree: **Illinski** (2013) claims that text cannot be considered a data type; **Šilić** (2010) argues that “unstructured text is not suitable for visualization”. Yet, as discussed above, most text visualizations transform the initial “unstructured” textual data into a reduced, structured dataset. This new dataset is no longer one-dimensional, but rather it constitutes a categorical or a network dataset and it can be represented with a wide range of tools that are not specific to text representation (**Hearst**, 2009; **Grobelnik; Mladenić**, 2002).

As we show in the cases we review here, most text visualizations do not represent raw data: that is, the text as it is. Rather what they do is transform the text into smaller chunks of data, normally extracting a representative part of that text. This process is one of data transformation and it occurs, for example, when a text is reduced to a list of words based on their frequency of appearance. In that case, the method chosen to represent the data will belong to a family of methods best suited to the data type. In this review we consider the most frequently employed strategies to represent single texts or collections of texts, paying special attention to strategies for representing textual data as it is, as a regular text, with all its complexities, irregularities and rich abstractions.

Text analysis is a key field for text visualization. Below, we present a brief commentary on this matter and its relationship with text visualization.

1.2. Text analysis

Text analysis, roughly synonymous with text mining (**Feldman; Sanger**, 2006), is an interdisciplinary field that includes information retrieval, data mining, machine learning, statistics, linguistics and natural language processing. According to **Marti Hearst** (2003), the goal of text mining is to discover “heretofore unknown information, something that no one yet knows and so could not have yet written down”. Text mining is a subfield of data mining whose typical applications include the analysis or comparison of literary texts, the analysis of biological and genomic data sequences and, more recently, the identification of consumer behaviour patterns or the detection of the fraudulent use of credit cards. **Hearst** differentiates these applications from information extraction operations, such as the extraction of people’s names, addresses or job skills. This latter task can be done with >80% accuracy, but the former, the full interpretation of natural language by a computer program, looks like it will not be possible for “a very long time” (**Hearst**, 2003).

To study text visualization and exploration it is important to examine the literature dedicated to both data visualization and text analysis, given the significant interrelationships that exist. Thus, while the text analysis output may limit the possibilities of visual presentation and interaction with the text, there is strong empirical evidence indicating that people learn better with a combination of text and illustration (visualization) than with text alone (**Anglin et al.**, 2004; **Levie; Lentz**, 1982).

2. Review

In this section we propose a possible classification based on the visual features that characterize the approaches to textual visualization and exploration, as identified in 49 cases.

The methodology to collect the cases is a two-part process. First, a traditional literature search and review (including practical examples and visualisation studies); and second, a subset of these have been selected, based on a preliminary analysis of their features. The aim was to select cases that provided a representative overview of the range of work in the field.

The classification of the cases is the product of empirical observation following an inductive analysis. The classification is followed by an analysis of these cases.

There are alternatives to those used in this paper for the selection and categorization of primary source methodologies such as **Kitchenham** (2004) and **Benavides; Segura; Ruiz-Cortés** (2010).

2.1. Classification of approaches

The basic classification of text visualization approaches comprises two categories according to the type of data to which they are applied:

1) Textual documents: that is, representations of single texts, where text is understood as a sequence of words ordered according to the hierarchy: document > paragraphs

> sentences > other punctuation marks > words > syllables and phonemes or morphemes. Where a text is a book or another kind of structure, then, it may have more granularities, including: chapters > sections > sub sections > etc. We also include the metadata of the text and other attached texts, i.e., title, author(s), publisher, copyright notes, acknowledgement, dedication, preface, table of contents, forward, glossary, bibliography, index, etc.

2) Text collections: that is, a group of texts in which each item constitutes a clearly differentiable entity. Typically when speaking of collections of texts, we speak of texts that have elements in common, be it their register, length or structure. All the cases we review here are collections of the same text type. Heterogeneous collections of texts are also referenced in the literature (Meeks, 2011), especially in representative analyses of a field of knowledge, where the aim is to include the greatest possible variety of expressions and vocabulary. In such cases the dataset can be said to be heterogeneous in term of its structure and register.

To these two data types, we then add several subjective subdivisions to each category according to the visual features used to represent the textual features. The aim here is to be able to describe and explain the cases under review, as well as to identify the key features of the text visualization approaches.

Single texts

- Whole <-> Part
- Sequential <-> Non sequential
- Discourse structure <-> Syntactic structure
- Search
- Time

Text collections

- Items <-> Aggregations
- Landscape
- Search
- Time

2.1.1. Single texts

In the specific instance of single texts, we classify the cases according to the part of the text that is represented, whether the approach follows the same sequence as that of the text, and the text structure employed in each case.

Whole or part?

In some instances, one part of the text is considered the essence of the text and is used in the visualization process rather than the whole text. Yet, there are processes that use the whole text, at least implicitly. Examples include:

- chapters of a book but not the whole text.
- representation of all the sentences of the text as coloured lines.
- verbs of a text, providing an impression of the style of the text.
- characters of a novel and their appearance within the text.
- places or dates present in the text.
- etc.

The cases in which the whole text is explicitly represented are, for obvious reasons, cases involving relatively short texts, e.g., song lyrics, speeches, poems, etc.

In some instances, such as when using *Radial word connections* (see, case 1 below) only certain words from the text are represented; yet, we classify this case as a whole text representation because the whole novel, chapter by chapter, is implicitly represented in the circle.

In those instances in which the whole text is represented (even implicitly) as one central element in the visualization, we classify it as being a whole-text visualization.

Does the visualization follow the same sequence as that of the text?

If the visualization follows the same sequence, or order, as that of the text, then the case is considered sequential; if not, then it is considered non-sequential. For example, a typical case that does not follow the same sequence as that of the original text would be a word cloud (see figure 1).

“Most text visualizations transform the initial ‘unstructured’ textual data into a reduced structured dataset”

Does the visualization use elements from discourse structure or from syntactic structure?

A text may present one of two kinds of structure that we consider useful for our research. One is so-called discourse structure. Depending on the nature of the text, the discourse structure can be completely subjective to the author's point of view—as in literature—or, restricted to a given structure—as in legal and scientific texts. In linguistics, discourse is a broad concept, but here we use it to refer to the parts of a text and the outline of a document: parts, chapters, sections, subsections, etc. The discourse structure is widely used when visualizing texts because it is a relatively straightforward way to represent the text sequence.

The second structure is the text's syntactic structure, referred to text structure in sentences, phrases and word classes—including verbs and nouns. This is an objective structure and is dependent on the rules of linguistics. In text visualizations, the elements comprising this structure, such as sentences, are very common.

2.1.2. Text collections

In the specific instance of text collections we classify the cases according to pure items or aggregations, i.e., as pure data or data landscapes. Thus we determine whether the items making up the collection can be differentiated or represented as aggregations. The specific questions we address are: How is each item in the collection graphically represented? Is each text represented as a graphical entity, i.e., as a point, a word or short sentence? Can the items in the visualization be counted, i.e., are they visually differentiated?

There are cases in which each item is not represented by a graphically distinct entity, but rather, for example, as a coloured block. Alternatively, the items are accumulated and shown as frequency distributions. When the items of the collection are not graphically distinct (visually countable)

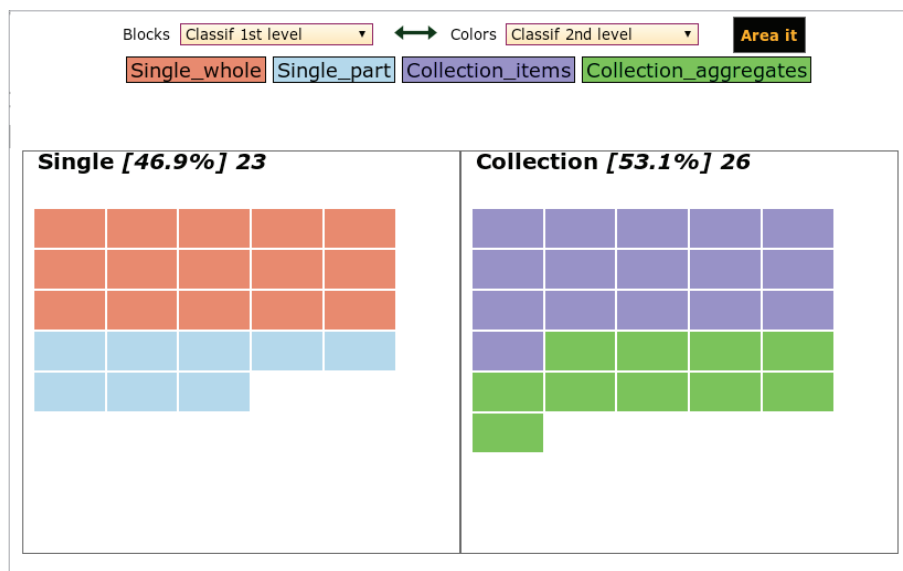


Figure 2. The 49 reviewed cases visualized with the Area software (screen shot).

then we speak in terms of the visualization of an aggregation rather than that of an item.

Pure data or data and landscape?

Are the items of the collection accompanied by any graphical content? Is another dataset, apart from that emanating from the text, also being represented? Some cases present the items embedded in a graphical environment, such as a map. This context might be an actual geographical map, a metaphor, or, for example, a conceptual landscape composed of words that form a second layer complementing that of the data collection, in which every distance plays a role: item-item (similarity between documents), word-item (importance of a word in a document), word-word (similarity between words in the collection).

Scales and axes are not considered as landscapes, nor are the elements of the interface in which the representation is embedded. This data layer, if not considered as the main dataset, would reduce substantially Tufte's data-ink ratio (Tufte; Graves-Morris, 1983) compared to the ratio of a pure data representation.

2.1.3. Both single texts and text collections

Properties that are equally applicable to single-text and text-collection visualizations include time, search results and dataset size.

Does time play a role?

Do the texts change over time? One set of visualization approaches highlights the changes undergone by a dataset over time. The most common approaches of this kind have been developed in computer science to represent code evolutions or in *Wikipedia* to indicate various aspects of article revisions.

This category also includes visualizations in which the dataset itself changes over time; for example, the visualization of the latest news will see the dataset grow over time.

Does the visualization result from a search query?

Visualizations of the output of information system retrieval is a well-defined kind of visualization characterized by the changing number of represented items depending on the number of search results obtained. This is a growing visualization subfield related to the disciplines of information systems and information retrieval (Mann, 2002; Hearst, 2009).

Validity for small or large datasets

It is rare that a visualization tool is independent of the size of the dataset that is to

be represented. Here, in those cases in which the tool has been clearly designed for a specific dataset size, the reader will be given the corresponding explanation.

2.2. Analysis of visualization approaches

We review a total of 49 cases applying the classification outlined above. In an attempt to incorporate the most crucial aspects of text visualization, our review concentrates on the specific ideas underpinning the text visualization, rather than the dataset and the contexts of each case.

Sixteen fields have been collected for each case: name, short name, author(s), year of publication, URL for further information, original dataset, discipline related to the work, description of the visualization method, description of the case, screen shot, thumbnail, classification (single or collection), classification (single-whole, single-part, collection-items, collection-aggregations), classification (time), classification (search), classification (dataset small, dataset large, N/A).

The cases are grouped into two sections and four subsections:

Single-text visualizations (23 cases)

- Whole-text visualizations (15 cases)
- Partial-text visualizations (8 cases)

Text collection visualizations (26 cases)

- Collection of items (16 cases)
- Collection of aggregations (10 cases)

For each subsection the cases are sorted by year of publication (descendant). To assist the reader, the collection of all reviewed cases can be viewed using the visualization and exploration software (also included in the review) known as AREA (Nualart, 2013).

2.2.1 Single-text visualization

We present single texts grouped as whole-text visualizations, partial-text visualizations and other subcategories.

The latter includes sequential and non-sequential visualizations, discourse-structures and syntactic-structures visualizations, search results and datasets dependent on time visualizations. Each subsection adheres to the following structure: list of cases, description of the group and discussion.

a) Whole-text visualizations

- 1) Literature. *Novel views: Les misérables*, *Radial word connections* by Jeff Clark (2013)
- 2) Literature. *Novel views: Les misérables*, *Character mentions* by Jeff Clark (2013)
- 3) Literature. *Poem viewer* by Katharine Coles et al. (2013)
- 4) Politics. *State of the Union 2011*, *Sentence bar diagrams* by Jeff Clark (2011)
- 5) Literature. *Visualizing lexical novelty in literature* by Matthew Hurst (2011)
- 6) Science/papers. *On the origin of species: The preservation of favoured traces* by Ben Fry (2009)
- 7) Science/papers. *Texty* by Jaume Nualart (2008)
- 8) Religion. *Bible cross-references* by Chris Harrison (2008)
- 9) Literature. *Literature fingerprint* by Daniel A. Keim and Daniela Oelke (2007)
- 10) Wikipedia. *History flow* by Fernanda Viégas and Martin Wattenberg (2003)
- 11) Literature. *Colour-coded chronological sequencing* by Joel Deshayé and Peter Stoicheff (2003)
- 12) Literature. *2-D display of time in the novel* by Joel Deshayé (2003)
- 13) Literature. *3-D display of time in the novel* by Joel Deshayé (2003)
- 14) Any. *Wattenberg's arc diagram* by Martin Wattenberg (2002)
- 15) Health. *TileBars* by Marti A. Hearst (1995)

Description

- Number of cases: We identify 15 cases that can be categorized as whole-text visualizations.
- Years: The cases were published over an 18-year period from 1995 to 2013.
- Authors: All the authors work in academic fields. The most prolific authors in this category are Jeff Clark and Joel Deshayé (with three cases each), followed by Martin Wattenberg (with two cases).
- Datasets: Most of the text corpora in this category are taken from literature (eight cases). Most authors draw on novels, especially well-known texts such as the classics, to demonstrate new visualization approaches.
- Methods: All the cases except case 14 (*arc diagram*) use colour as part of the visualization method. Five cases use

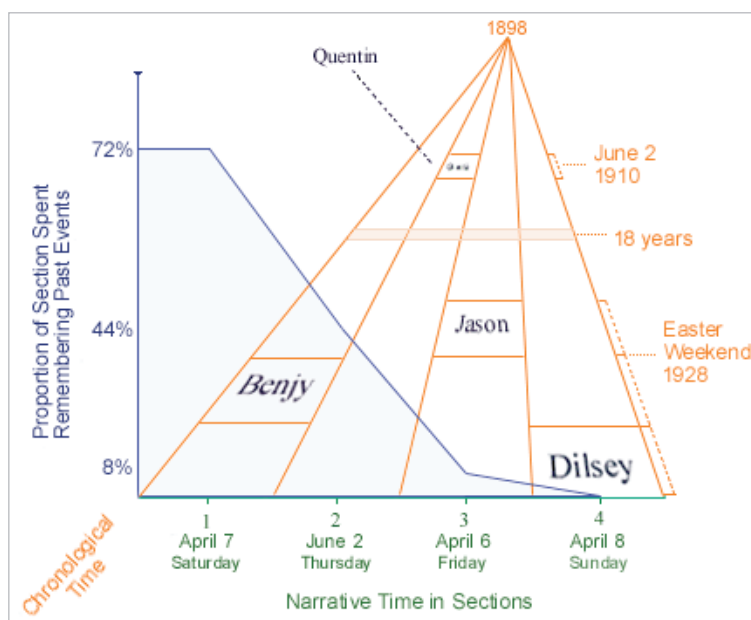


Figure 3. (Case 13) 3-D display of time of William Faulkner's novel *The Sound and the Fury*, by Joel Deshayé and Peter Stoicheff (2003)

methods that are bar chart derivatives (cases 4, 5, 6, 9 and 11). Three cases use curves connecting parts of the texts: two arcs and one radial diagrams (cases 1, 8 and 14).

Discussion

A common method cannot be identified for these whole-text visualizations. Yet, as expected, they all present an axis representing the whole text. In 13 of the 15 cases, the text line is represented by a horizontal or vertical line. The two exceptions use a circle—the case of *Radial word connections* (case 1)—and an iconification of a text on the page—the case of *Texty* (case 7).

Since whole-text visualizations always include an abstraction of the text, referred to as its text line, a question arises: which part of the text is physically present in the whole-text visualization being reviewed? Interestingly, nine of the 15 visualizations do not show a single word (cases 4, 5, 6, 7, 8, 9, 10, 11 and 15). Four cases show a small number of words (cases 1, 2, 12 and 13) (figure 3), while only two cases show all the text (cases 3 and 14).

The most common approach is to show the occurrence of a certain feature—this might be a term, topic, cross-reference or character—within the text as a whole (all cases except 3, 12, 13 and 15). With the exception of Wattenberg's *arc diagrams* (case 14), these occurrences are represented using the same colour.

It is interesting to observe how very similar data are represented in very different ways depending on the case under review. For example, while Viégas and Wattenberg's *History flow* (case 10) and Fry's *Favoured Traces* (case 6) both present document-version histories by section, the former is spatialized and the latter animated. Similarities, however, are seen in the approaches adopted, for example, by *TileBars* (case 15) and *Texty* (case 7). Thus, both highlight words from the text within a rectangular figure that is representa-

tive of the whole text. Other cases use opposite or complementary techniques. Thus, Wattenberg's *Arc diagram* (case 14) shows repetitions while Hurst's novelty visualization (case 5) shows only new strings, and no repetitions.

Literature and other complex texts, such as political speeches (case 4) and the *Bible* (case 8), dominate the type of corpora used in this category (10 cases). This is perhaps surprising, as these texts tend to be complex, often presenting a high level of abstraction and little formal structure. Arguably, when opting to introduce or test a new approach, it would make more sense to work with simpler, more structured texts (such as scientific papers, patents, health diagnostics, etc.) that present greater regularity in terms of their vocabulary, text length, discourse structure and register. Given the inherent freedoms associated with literature, novelists are under no obligation to adhere to any pattern or rule that might help us give structure to the unstructured.

However, depending on how the text is treated and processed, the nature of the text is not always relevant. For example, Matthew Hurst (case 5) tracks the introduction of new terms in literary texts. Yet the tool can be applied to any other text type, its results being unrelated to the complexity of the text given the ubiquity of the method. Having said this, it would be interesting to apply the technique to scientific papers in which the style is much more clearly defined. Similar arguments can be applied to *Radial word connections* (case 1), *Sentence bar diagrams* (case 4) and *Literature fingerprints* (case 9).

b) Partial-text visualizations

16) Literature. *Novel views: Les misérables. Characteristic verbs* by Jeff Clark (2013)

17) Any. *Wordle* by Jonathan Feinberg (2009)

18) Books. *DocuBurst* by C. Collins, S. Carpendale and G. Penn (2009)

19) Literature. *Phrase nets* by Frank van Ham, Martin Wattenberg and Fernanda B. Viégas (2009)

20) Google data. *Word spectrum: Visualizing Google's bi-gram data* by Chris Harrison (2008)

21) Google data. *Word associations: Visualizing Google's bi-gram data* by Chris Harrison (2008)

22) Literature/songs. *Document arc diagrams* by Jeff Clark (2007)

23) Any book. *Gist icons* by P. DeCamp, A. Frid-Jimenez, J. Guinness, D. Roy (2005)

Description

- Number of cases: We identify eight cases that can be categorized as partial-text visualizations.
- Years: The cases were published over an eight-year period from 1995 to 2013.
- Authors and datasets: Two cases by Jeff Clark (cases 16 and 22) and one by the creative team of Wattenberg and Viégas in collaboration with van Ham (case 19) use literary texts. The two cases by Chris Harrison use large *bi-gram* datasets published by Google. One case is not dependent on the nature of the text: *Wordle* (case 17), the very popular "word cloud" method introduced by Feinberg. Finally, two interactive approaches involving large datasets are presented: *DocuBurst* (case 18) and *Gist icons* (case 23).
- Methods: In six of the eight cases (cases 16, 17, 18, 19, 22 and 23), the dataset is reduced to what is called a bag of words and only these words are present in the visualization. Cases 20 and 21 are representations of all bi-grams that pit two primary terms against each other.

Discussion

Partial-text visualization is a successful, popular way to draw a text, presumably because of the way in which a long text can be effectively represented using a small set of words. Simple statistical methods, such as word frequency counts, are readily interpretable. A list of variously sized words is a direct way of communicating with any user, from beginner to expert. Most of the partial-text approaches available online use statistical methods to extract the part from the whole.

It is our contention that extracting part of the corpora can be affected by the structure and complexity of the whole. In the visualizations under review, half present unstructured text corpora, but the criteria used in extracting the part from the whole are well defined and include lists of verbs (*Characteristic verbs*, case 16), words occurring in the text in an "X and Y" pattern (*DocuBurst*, case 18) and lists of words not included in a list of predefined empty words (*Google's bi-gram data*, case 21).

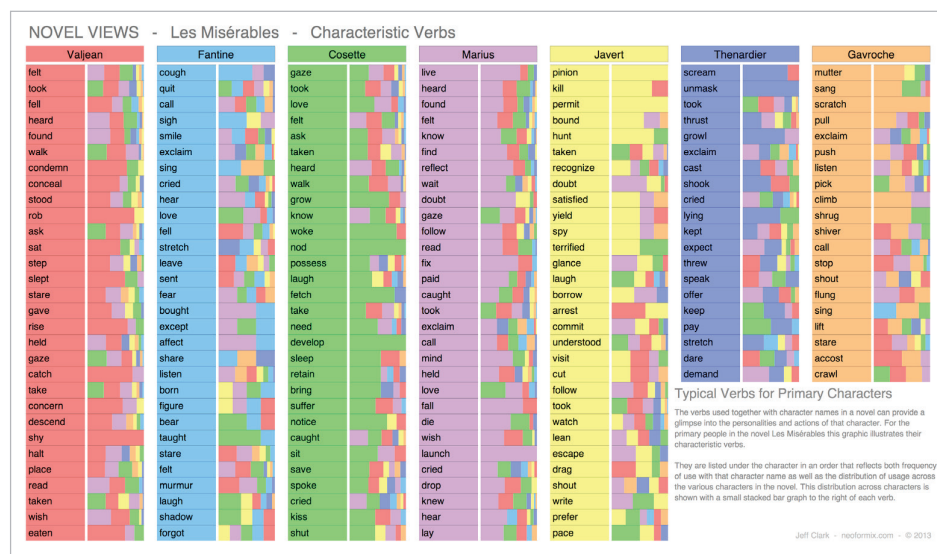


Figure 4. (Case 16) *Novel views: Les misérables. Characteristic verbs* by Jeff Clark (2013)

Clearly, extraction processes based on word or phrase functionality, as opposed to those that use statistical methods, are more closely affected by the nature of the text. Here, we focus on these cases because they are more interesting in terms of our research goals. They include the cases of *Novel views: Les misérables. Characteristic verbs* (case 16), which represents only verbs, *DocuBurst* (case 18) which uses the crowd-sourced lexical database *Wordnet* as a human-like backup, and *Phrase net* (case 19) and the two *Google bi-gram visualizations* (cases 20 and 21).

A common pattern detected in the partial-text visualizations reviewed is that once a part of the text has been extracted all except one (*Document arc diagrams*, case 22) discard any reference to the original text sequence in the visualization. See the following point for a more detailed discussion of this idea.

c) Other subcategories

Here we include sequential and non-sequential visualizations, discourse and syntactic structures visualizations, search results and datasets dependent on time visualizations.

Sequential visualizations

Sixteen of the 23 single-text visualizations maintain a similar sequence to that of the original text. Seven of these visualize the sequence using a discourse structure (primarily chapters), while the remaining nine use syntactic elements to represent the original sequence of the text (primarily words).

Strikingly, only one partial-text visualization, Clark's *Document arc diagrams* (case 22) (figure 5), follows the original text sequence, whereas all the whole-text visualizations are sequential. It would thus appear that sequentiality is intrinsic to whole-text visualization. Whole-text visualizations do not literally represent every word of the text, but rather present a graphical metaphor of the whole: a text line. This text line may represent either a discourse structure or a syntactic structure of the text; but, whatever the case, graphically a line or area is used to represent the length of the text.

The sequentiality of the visualization means it can be read both backwards and forwards, as can the text. In the case of a long text, such as a book (nine of the 16 cases), the visualization can serve as a map or guide to the text.

Non-sequential visualizations

Five cases use non-sequential visualizations: three use word clouds (cases 17, 20 and 21), one a net of phrases (case 19) and one visualizes all the verbs in the text (case 16).

Discourse structures in the visualization

Cases: 1, 2, 5, 6, 8, 11, 12 and 13

The eight visualizations that follow the discourse structure of the text are sequential –no cases being found in which

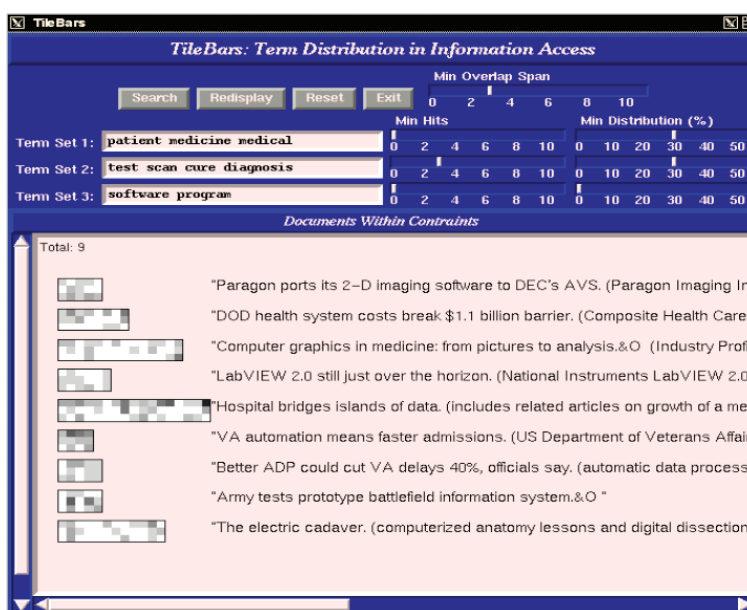


Figure 5. (Case 15) *TileBar* search on (patient medicine medical AND test scan cure diagnosis AND software program) with stricter distribution constraints.

the discourse structure appeared out of sequence with regards to the text. This is perhaps unsurprising, as those cases in which the text is divided into chapters and each chapter represented as a separate entity were considered as text collection visualizations (e.g., *Sentence bar diagrams*, case 4). For this reason, all the cases in this section represent the parts of a text ordered and aligned (in a curve or line). Of the eight visualizations, five represent chapters or sections of a book, two represent complete volumes, while one (*Colour-coded chronological sequencing*, case 11) divides the text in colours according to narrative topics and scenes. Indeed, case 11 is the only one we have identified that uses discourse structure elements that are more deeply embedded than chapters, sections, books and volumes. In all likelihood, more deeply embedded methods than these, such as, narrative topics, would require manual text line segmentation.

Syntactic structures in the visualization

Cases: 3, 16, 4, 7, 18, 9, 22 and 23.

The other eight sequential visualizations use intrinsic text elements, including groups of words (cases 7, 18, 22 and 23), verbs (case 16), sentences (cases 4 and 9) and a complete text analysis (case 3). Syntactic analysis requires either word-by-word parsing of the text (using a database of lexical or semantic word lists) or sentence and paragraph parsing. Syntactic-structure visualization is less dependent on the nature of the text in the sense that the methodology is unaffected by the complexity of the text. Typically, the software automatically extracts or marks the chosen syntactic elements.

Search-result visualizations

Cases: 15, 18 and 23

The three search-result visualizations were presented as web applications and were, therefore, interactive – the user being able to query the visualization system and obtain a

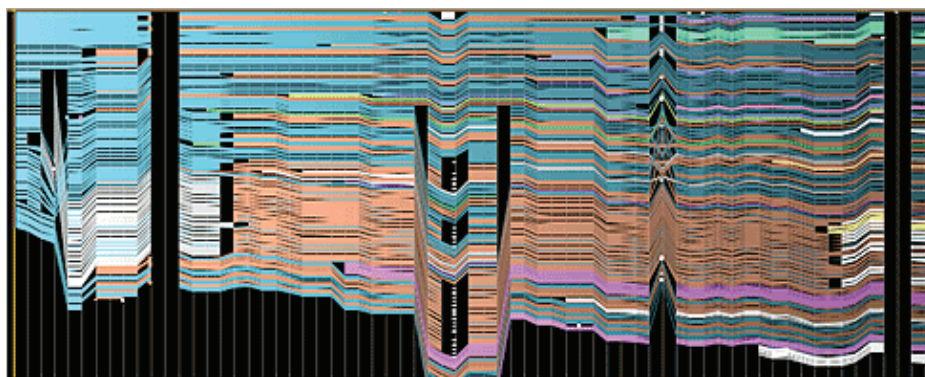


Figure 6. (Case 10) *History flow* by Fernanda Viégas and Martin Wattenberg researchers at IBM's Visual Communication Lab (2003)

over time. A dynamic text visualization demonstrates that data visualization may be the only way to solve certain tasks and that it is not just one more method of pure data advocacy. For example, it is extremely challenging to show how a *Wikipedia* entry evolves over time in line with the editors' participation (*History flow*, case 10) (figure 6). *History flow* provides a solution to

unique representation for each search. The three cases, however, are no longer available online. *DocuBurst* (case 18) is a *Prefuse* application that can be downloaded (Collins et al., 2009). *Prefuse* is a set of software tools for creating rich interactive data visualizations.

TileBars is a classic case of visualization (cited 625 times by *Google Scholar*) designed by a leading expert in visualization and search engine interfaces, Marti Hearst. *DocuBurst* and *Gist icon* are interactive radial visualizations, the latter being one of the references and main influences on the development of *DocuBurst*, as explained in the *DocuBurst* paper cited.

Partial-text visualization is a successful, popular way to draw a text, presumably because of the way in which a long text can be effectively represented using a small set of words

Search-result visualization approaches have not been widely implemented in information retrieval systems and most result outputs are one-dimensional lists of itemized texts (Nualart; Pérez-Montoro, 2013). The three cases reviewed here are each applied to large datasets and, starting with a search query, present an improved search output designed to help the user read and filter the results. All three are particularly concerned with distinguishing between similar items: *TileBars* searches *PubMed* (more than 20 million papers); *DocuBurst* uses the *WordNet* lexical database (155,287 words organized in 117,659 synsets for a total of 206,941 word-sense pairs) to classify the visualized text; and, *Gist icons* use, among others, the complete dataset of approximately 7 million *USpto* patents and the *Enron* email dataset comprising 500,000 emails.

In the text collection category below, we present nine further search-result visualizations.

Time dependent datasets

Cases: 6 and 10.

We present two cases in which the visualization approaches can be used to understand or follow the evolution of a text

this problem and sheds light on the complex collaborative process of *Wikipedia*.

In the second case (*Favoured traces*, case 6), an animated visualization demonstrates how Darwin's ideas evolved through successive editions of the *Origin of Species*. In Ben Fry's words: "The first English edition was approximately 150,000 words and the sixth is a much larger 190,000 words. In the changes are refinements and shifts in ideas—whether increasing the weight of a statement, adding details, or even a change in the idea itself."

2.2.2. Text collections

We present text collections grouped as pure item visualizations, aggregation visualizations and other subcategories. The latter includes data as a landscape layer and search result visualizations. Each subsection adheres to the following structure: list of cases, description of the group and discussion.

a) Item visualizations

24) Literature (Note: this converts a single text into a collection). *Novel views: Les misérables. Segment word clouds* by Jeff Clark (2013)

25) Literature. *Grimm's fairy tale network* by Jeff Clark (2013)

26) Twitter. *Spot* by Jeff Clark (2012)

27) Science. *Word storm* by Quim Castella and Charles Sutton (2012)

28) Literature. *Topic networks in Proust. Topology* by Elijah Meeks and Jeff Drouin (2011)

29) Wikipedia. *Notabilia* by D. Taraborelli, G. L. Ciampaglia and M. Stefaner (2010)

30) Media art. *X by Y* by Moritz Stefaner (2009)

31) Search engine. *Search clock* by Chris Harrison (2008)

32) Online media. *Digg rings* by Chris Harrison (2008)

33) Science. *Royal Society Archive* by Chris Harrison (2008)

34) Wikipedia. *WikiViz: Visualizing Wikipedia* by Chris Harrison (2007)

35) Visualization. *Area* by Jaume Nualart (2007)

36) *Chromograms* by M. Wattenberg, F.B. Viégas and K. Hollenbach (2004)

37) Search engines. *KartOO/Ujiko* by Laurent Baleyrier and Nicholas Baleyrier (2001)

38) Search engines. *Touchgraph* by TouchGraph, LLC. (2001)

39) Internet. *HotSauce* by Ramathan V. Guha (1996)

Description

- Number of cases: We identify 16 cases that can be categorized as item visualizations.
- Years: The cases were published over a 17-year period from 1996 to 2013.
- Authors: The most prolific authors in this category are Chris Harrison (cases 13, 32, 33 and 34) and Jeff Clark (cases 24, 25 and 26), followed by Moritz Stefaner with two cases (29 and 30).
- Disciplines and datasets: Interestingly, nine cases are datasets taken from the Internet: *Wikipedia* (cases 29, 34 and 36), search engines (cases 31, 37 and 38), *Twitter* (case 26), online media (case 32), web pages (case 39). Only three cases use literary texts (cases 24, 25 and 28). Finally, two cases visualize scientific papers (cases 27 and 33), one case uses media art datasets (case 30) and one represents non-specific collections (case 35).

Discussion

The main difference between single-text and text-collection visualizations lies in the nature of the text. In the case of the latter, most of the texts do not originate from literature and are accessible online. Yet, the nature of the text appears to be less important when the goal is the representation of the collection rather than of the text itself.

Item visualizations use methods that are independent of the nature of the items themselves. Once the text collections have been itemized, the dataset can be considered a general case of data visualization and not a pure case of text visualization. For this reason, in this category, the methods are generally well known and used in other fields of visualization. Thus, we find six network visualizations (cases 25, 28, 34, 37, 38 and 39), three timelines (cases 31, 32 and 33) and three cases that likewise use timelines but which also permit categorization-based groupings (cases 26, 30 and 35) (figure 7).

Finally, four cases are, we believe, quite specific to text visualization. Two are concerned with item com-

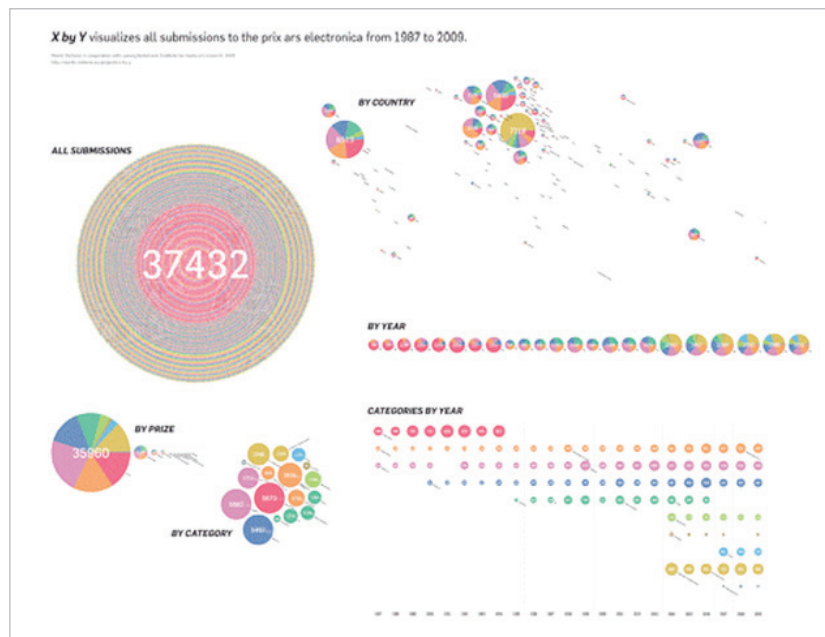


Figure 7. (Case 30) *X by Y* by Moritz Stefaner (2009)

parison: *Segment word clouds* (case 24) and *Word storm* (case 27). *Segment word clouds* transforms a single text into a text collection. Specifically, it is used to represent the chapters of *Les misérables* as word cloud items, thus facilitating their comparison. It also uses colour to identify words as they acquire prominence in the text.

Word storm is a reinvention of word cloud, or more specifically a variation of *Wordle* (case 17) that allows word clouds to be compared. This is achieved by assigning a fixed position to each word. This simple idea makes it visually easy to compare word clouds while maintaining the usual word cloud features.



Figure 8. (Case 29) *Notabilia. 100 longest Article for deletion [AfD] discussions on Wikipedia* by Dario Taraborelli, Giovanni-Luca Ciampaglia (data and analysis) and Moritz Stefaner (visualization) (2010)



Figure 9. (Case 43) *Web seer* by Fernanda Viégas & Martin Wattenberg (2009)

To conclude, *Notabilia* (case 29) and *Chromograms* (case 36) are two highly original cases that deserve mention. The very specific design of *Notabilia* shows the evolution of “Article for deletion” discussions of Wikipedians (figure 8), discussions that are sometimes more like “flame wars” given the controversies that rage over the simple existence of certain definitions. *Notabilia* visualizes the evolution of the hundred longest discussions and their final outcomes. Moritz Stefaner’s visualization constitutes an interactive bushtree, the branches of which are highlighted when moused over. The shape of the branches informs the reader about the nature of the discussion: cyclical, straight or never-ending.

Chromograms is also based on *Wikipedia* data, providing an analysis of the comments of editors for each edition of a *Wikipedia* entry. Visually it produces colour-coded stripes that in a small space rapidly inform the reader about the edit history of *Wikipedia* entries.

“ It might prove more effective to apply visualization techniques to texts that have a more formal register and/or predefined outline and a well-defined vocabulary ”

b) Aggregation visualizations

- 40) Literature. *Grimm’s fairy tale metrics* by Jeff Clark (2013)
- 41) Topic models. *Termite* by J. Chuang, C.D. Manning and J. Heer (2012)
- 42) *Wikipedia*. *Pediameter* by Müller-Birn, Benedix and Hantke (2011)
- 43) *Google* suggestions. *Web Seer* by Fernanda Viégas & Martin Wattenberg (2009)
- 44) *Google* n-grams. *Web trigrams: visualizing Google’s trigram data* by Chris Harrison (2008)

45) Political speech. *Feature-Lens* by A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. Clement, B. Shneiderman and C. Plaisant (2007)

46) Online news. *Newsmap* by Marcos Weskamp (2004)

47) Email conversation. *Themail* by Fernanda B. Viégas, Scott Golder, Judith Donath (2006)

48) Search engine. *WebBook* by S.K. Card, G.G. Robertson and W. York (1996)

49) Any texts. *Dotplot applications* by Jonathan Helfman (1994)

Description

- Number of cases: We identify 10 cases that can be categorized as aggregation visualizations.
- Years: The cases were published over a 19-year period from 1994 to 2013.
- Authors and datasets: Only Fernanda B. Viégas participated in more than one of the 10 cases in this category (cases 43 and 47); the rest participated in just one case each. The texts are very similar in nature to those in the item visualization category. Five cases are corpora that can be found online (*Wikipedia*, case 42; *Google*, cases 43 (figure 9) and 44; online news, case 46; search engine results, case 48). The standard unstructured texts include one from literature (*Sentence Bar Diagrams*, case 4), one from political speeches (*FeatureLens*, case 45) and one from a year’s worth of email conversations between two correspondents (*Themail*, case 47). Finally, there are two quite unique cases: *Termite* (case 41) and *Dotplot* (case 49). All the cases are discussed below.

Discussion

Aggregation visualizations is the category with the greatest variation in the methods employed. Thus, apart from visualizing text collections, the only thing the 10 cases assigned to this category have in common is that they do not represent specific items.

Given these circumstances, we comment on each case separately:

Sentence bar diagrams (case 40) provide a matrix (or table-like) visualization that allows rows to be sorted by clicking on columns. The columns provide a quantitative definition of 13 metrics related to the 62 stories making up Grimm’s fairy tales. It is a powerful tool for analysing, understanding and comparing the tales.

Termite (case 41) is a case that represents an intermediary dataset known as topic models. Topic models are a “cleverer” way of obtaining a bag-of-words from a text than applying a typical word-frequency statistical analysis. *Termite* does not visualize texts but it does compare parts of

texts. As such, the tool can be used to compare topic models.

Pediameter (case 42) is a specific interface that uses bar charts to show *Wikipedia* editions in real time. It is most remarkable for using a device known as an *Arduino* to detect editions and transcribe them to a physical indicator, merging digital and material worlds.

Web Seer (case 43) is another specific visualization method that shows the most popular search queries based on *Google* suggestions. The approach allows queries to be compared by representing the suggestions with trees and then connecting the matching branches. The simplicity of this case contrasts with its power of communication: rapid and user friendly.

Google's tri-gram data (case 44) uses a similar visualization method to that used by *Web seer*. It draws on the huge *Google* n-gram dataset and represents and compares three-word sentences (tri-grams).

FeatureLens (case 45) is an interactive, dashboard-style interface for comparing texts. The central representation uses a visualization of frequent concepts similar to that used by *Texty* (case 7) and *TileBars* (case 15). It allows text browsing and shows line graphs of frequent words found throughout a text.

Newsmap (case 46) uses treemap visualization to offer a new method for reading and monitoring the news in real-time, employing online *Google* news feeds. It is totally customizable in terms of topic, country and publication time. The software, which is available free of charge online, can also be used for news searches.

TheMail (case 47) is an experiment in which a highly specific interface was developed to follow and analyse the evolution of an email correspondence between two people over the course of one year. It visualizes the words that characterize each of the writers and their evolution over time.

When first developed in 1996, *WebBook* (case 48) (figure 10) was a somewhat surprising application, as it transformed search engine results in a multimedia (text and images, primarily) mash-up based on the metaphor of the book. The application was a pure text (web pages) collection visualization that presented the results as aggregations of text and images.

Finally, *Dotplot* (case 49) was an innovative visualization application with multiple uses, not unlike *Arc diagrams* (case 14). The main use of *Dotplots* is for text comparisons, including multi-language, text version and programming code comparisons.

c) Other subcategories

Here we include landscape data layers, search-result visualizations and time-dependent datasets.

Landscape as an additional data layer

Cases: 40, 26, 28, 33, 47, 37, 38 and 49.



Figure 10. (Case 48) *WebBook* by Stuart K. Card, George G. Robertson, and William York (1996)

The typical concept of landscape data is a network visualization comprising two layers of data, as in *Topic networks* (case 28). In this specific case, the first layer is provided by the Marcel Proust texts represented as items and the second layer by a network of topic models of these texts. The positions of the nodes of both layers are optimised so that proximity indicates more strongly related nodes. This definition of landscape can also be found in the defunct search engine results provided by *KartOO/Ujiko* (case 37) and *TouchGraph* (case 38).

All the other cases included in this category present text collections in combination with more data. This is the case of *Dotplot*, which represents the coincidence or otherwise of strings in various texts, and of *Grimm's fairy tale metrics*, which combines a list of texts in rows with various parameters listed in columns. These parameters do not form a direct part of the text, but rather they are recalculated features related to the text, including, for example, length, lexical diversity and the presence of different groups of words that represent entities (for example: body -> hand, head, heart, eyes and foot) in each tale.

A third kind of landscape is based on the representation of timed metadata, as exemplified by *Spot* (case 26), the *Royal Society Archive* (case 33) and *TheMail* (case 47).

A common feature of landscape visualizations is their capacity to compare a collection of texts simultaneously with a second parameter, while their main limitation is the number of items represented so that large numbers create problems of overlapping items.

Search result visualizations

Cases: 26, 43, 35, 45, 47, 46, 37, 38 and 48.

Compared to single-text visualizations, text-collection visualizations include considerably more cases offering search capacities (three vs. nine). Common sense suggests that when presenting a text collection, a natural feature of such an approach will be a way of selecting part of that collection based on given criteria, i.e., filter and search features.

All the cases included in this category allow search queries and output a unique visualization for each query. All the cases include a search box and a search button.

Time-dependent datasets

Cases: 42, 29, 36 and 46.

The four cases included in this category allow the user to monitor the evolution of the texts in the collection over time. Only one is designed for use in real-time (*Newsmap*, case 46), but potentially all of them can visualize the collection on a specific date and at a specific time.

One obstacle faced by an approach that represents changes in text collections over time is providing access to an updated feed or an accessible API. It is presumably for this reason that three of the four use *Wikipedia* data and the other uses *Google* news. In all cases, they are online sources that have long allowed public access to their feeds.

3. Conclusions

The diversity of approaches developed in different disciplines, the wide diffusion of publications or, on occasions, the absence of formal publications of innovative ideas, represent a considerable challenge to the undertaking of a comprehensive survey of the work completed in this field. Thus, some of the visualizations we present here have been unearthed in highly specific publications, the case for example of Joel Deshayé and Peter Stoicheff and their work on representing Faulkner (cases 11, 12 and 13). If we read Stoicheff's working notes it is apparent that their visualizations were developed to facilitate the study of William Faulkner's narrative timelines. There are no additional references to the application of these interesting ideas to other texts, suggesting that more works remain hidden in the depths of other fields.

Text visualization, as we have argued throughout this review, may be considered a subfield of data visualization. Yet, the boundaries of the discipline are not always clearly defined. This is readily illustrated, for example, by the case of Harrison's *Search clock* (case 31), in which the text corpora comprise an enormous dataset of search engine queries. Can this dataset really be considered a collection of texts when each of them, in most instances, is no more than one or two words in length? Does a text have to satisfy a minimum length in order to be considered a text? Here, we opted to treat case 31 as a collection of texts, short ones admittedly but, ultimately, *texts*.

Clearly, the critical decision to be made throughout this review has been how to classify the cases identified. As few papers have attempted to review only text visualization approaches, we turned to classic data visualization reviews (e.g., Shneiderman, 1996) as well as to more recent ones (e.g., Collins *et al.*, 2009). In all these instances, the classifications were based on tasks that the visualization approach can solve rather than on the explicit aspects of the visualization themselves. For this reason we chose to propose our own classification, which, while far from perfect, we hope will be useful for undertaking a classification based on visual features.

We conclude with a list of insights, as well as shortcomings, that we have identified to date:

- Single-text visualizations have been applied mainly to literature, a field that, apart from being characterized by complex combinations of words, can present high levels of human abstraction and freedom of structure and experimentation. As such it might prove more effective to apply visualization techniques to texts that have a more formal register and/or predefined outline and a well-defined vocabulary, such as legal texts, scientific papers, template-based texts and communications, etc.
- We have identified only one single/partial-text visualization that is sequential (*Document arc diagrams*, case 22). Most partial-text visualizations extract the essence of the text based on one or more criteria and so the original sequence of the text is lost. Since sequential visualization approaches present certain advantages, it seems that partial-visualization approaches that maintain the original text sequence should be encouraged.
- Text-collection visualizations tend to employ methods that are used for data visualization in general. Hence, there is a need for further experimentation in applying more standard data visualization methods and approaches to the specific subfield of text visualization.
- Text collection aggregations is the category in which the most specific designs and ideas have been developed. More work needs to be undertaken to identify any common approaches in this kind of visualization.

And, finally, we pose the following question:

- Why is it that most of the cases reviewed here that are more than five years old are no longer available online? If the software used is no longer (or was never) in use, we should perhaps question its effectiveness. While we have not investigated just how many cases form part of commercial software products and how many, following publication, have simply been forgotten, the question remains as to why some apparently magnificent ideas did not establish themselves as new standards. Our challenge to researchers is to produce applications that will be adopted in one field or another, or which can solve a problem for a certain group of users; indeed, as the cases reviewed here highlight, adoption seems to represent a considerable challenge.

Acknowledgement

This work is part of the project "Active audiences and journalism. Interactivity, web integration and findability of journalistic information". CSO2012-39518-C04-02. *National plan for R+D+i, Spanish Ministry of Economy and Competitiveness*.

4. References

- Anglin, Gary J.; Vaez, Hossein; Cunningham, Kathryn L. (2004). "Visual representations and learning: The role of static and animated graphics". *Handbook of research on educational communications and technology*, 2, pp. 865-916.
- Baeza-Yates, Ricardo; Ribeiro-Neto, Berthier *et al.* (1999). *Modern information retrieval*. New York: ACM press, vol. 463.

- Baeza-Yates, Ricardo; Broder, Andreiz; Maarek, Yoelle** (2011). "The new frontier of web search technology: Seven challenges". *Search computing*, v. 6585 of *Lecture notes in computer science*, pp. 3-9.
http://dx.doi.org/10.1007/978-3-642-19668-3_1
- Benavides, David; Segura, Sergio; Ruiz-Cortés, Antonio** (2010). "Automated analysis of feature models 20 years later: A literature review". *Information systems*, v. 35, n. 6, pp. 615-636.
<http://dx.doi.org/10.1016/j.is.2010.01.001>
- Collins, Christopher; Carpendale, Sheelagh; Penn, Gerald** (2009). "DocuBurst: Visualizing document content using language Structure". *Computer graphics forum* (Procs. of the *Eurographics/IEEE-VGTC Symposium on visualization*, EuroVis), v. 28, n. 3, pp. 1039-1046.
<http://dx.doi.org/10.1111/j.1467-8659.2009.01439.x>
- Feldman, Ronen; Sanger, James** (2006). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press. ISBN: 13 978 0 521 83657 9
- Grobelnik, Marko; Mladenić, Dunja** (2002). "Efficient visualization of large text corpora". In: *Procs of the 7th seminar*. Dubrovnik, Croatia.
<http://ailab.ijs.si/dunja/SiKDD2002/papers/GrobelnikSep02.pdf>
- Hearst, Marti A.** (2003). *What is text mining?*
<http://people.ischool.berkeley.edu/~hearst/text-mining.html>
- Hearst, Marti A.** (2009). "Search user interfaces", Chapter 1. ISBN: 9780521113793
<http://searchuserinterfaces.com/book>
http://searchuserinterfaces.com/book/sui_ch1_design.html
- Hearst, Marti A.** (2011). "Natural search user interfaces". *Communications of the ACM*, v., 54, n. 11, November, pp. 60-67.
<http://cacm.acm.org/magazines/2011/11/138216-natural-search-user-interfaces/fulltext>
<http://dx.doi.org/10.1145/2018396.2018414>
- Heer, Jeff** (2010). "A conversation with Jeff Heer, Martin Wattenberg, and Fernanda Viégas". *Queue*, v. 8, n. 3, 10 pp., March.
<http://doi.acm.org/10.1145/1737923.1744741>
- Iliinsky, Noah** (2013). *Choosing visual properties for successful visualizations*. IBM Software. Business Analytics.
<http://public.dhe.ibm.com/common/ssi/ecm/en/ytw03323usen/YTW03323USEN.PDF>
- Kitchenham, Barbara** (2004). *Procedures for performing systematic reviews*. Keele, UK, Keele University, 33 pp.
- Levie, W. Howard; Lentz, Richard** (1982). "Effects of text illustrations: A review of research". *ECTJ*, v. 30, n. 4, pp. 195-232.
- Mann, Thomas M.** (2002). *Visualization of search results from the world wide web*.
<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.91.2535>
- Meeks, Elijah** (2011). *Digital humanities specialist*. Documents.
<https://dhs.stanford.edu/comprehending-the-digital-humanities/documents>
- Nualart-Vilaplana, Jaume** (2013). *How we draw texts: a visualization of text visualization tools*.
<http://research.nualart.cat/textvistools>
- Nualart, Jaume; Pérez-Montoro, Mario** (2013). "Texty, a visualization tool to aid selection of texts from search outputs". *Information research*, v. 18, n. 2, June.
<http://www.informationr.net/ir/18-2/paper581.html>
- Shneiderman, Ben** (1996). "The eyes have it: A task by data type taxonomy for information visualizations". In: *Visual Languages*. Proceedings IEEE Symposium, pp. 336-343.
<http://dx.doi.org/10.1109/VL.1996.545307>
- Šilić, Artur; Dalbelo-Bašić, Bojana** (2010). "Visualization of text streams: A survey". *Knowledge-based and intelligent information and engineering systems*, v. 6277 of *Lecture notes in computer science*, pp. 31-43. Berlin, Heidelberg: Springer.
http://dx.doi.org/10.1007/978-3-642-15390-7_4
- Stefaner, Moritz** (2013). *Gender balance visualization*.
<http://moritz.stefaner.eu/projects/gender-balance/#NUM/NUM>
- Strecker, Jacqueline** (2012). *Data visualization in review: summary*. International Development Research Centre (IDRC), Ottawa, ON, Canada.
<http://idl-bnc.idrc.ca/dspace/bitstream/10625/49286/1/IDL-49286.pdf>
- Times Higher Education. *World university rankings 2012-2013*.
<http://www.timeshighereducation.co.uk/world-university-rankings/2012-13/world-ranking>
- Tufte, Edward R.; Graves-Morris, P. R.** (1983). *The visual display of quantitative information*, v. 2. Cheshire, CT: Graphics Press, 199 pp.

A.2. Paper II

Citation:

Nualart, Jaume; P´erez-Montoro, Mario (2013). Texty, a visualization tool to aid selection of texts from search outputs. *Information Research*, 18(2) paper 581.

Affiliation (in catalan)

- 1r autor: Jaume Nualart Vilaplana es doctorand a la Facultat de Biblioteconomia i Documentació de la Universitat de Barcelona, enginyer de recerca al NICTA (Australia), i doctorand a la Faculty of Arts and Design, University of Canberra (Australia).
- 2n autor: Mario Pérez-Montoro, doctor i professor al Dept. de Ciències de la Informació de la Universitat de Barcelona.

Summary (in Catalan)

Introducció: La presentació de la pàgina de resultats en un sistema de recerca té un paper important en la satisfacció de les necessitats d'informació d'un usuari. Els criteris de gestió del rendiment habituals i les eines per a organitzar aquests resultats tenen limitacions que poden dificultar la satisfacció dels usuaris. Presentem Texty com un nou enfocament que pot ajudar a millorar l'experiència de recerca dels usuaris.

Mètode: El corpus de textos als quals s'ha aplicat Texty són articles científics de la revista "Information Research". Per tal de filtrar els textos hem construït grups de paraules o vocabularis de cinc camps concrets de coneixement: àmbit conceptual, àmbit experimental, metodologia qualitativa, metodologia quantitativa i Computació/Programari.

Resultats: Mostrem com Texty, intrínsecament, és capaç de codificar o oferir informació sobre textos que altres representacions clàssiques (diagrames de barres i de línies, principalment) no són capaços d'oferir.

Conclusions: Texty és una eina complementària que millora la interacció intel·lectual amb llistats de documents textuais, la qual cosa permet als usuaris fer una tria de textos de manera més efectiva coneixent-ne la seva estructura abans de llegir-los.

Access online (URL)

- <http://InformationR.net/ir/18-2/paper581.html>

- [Contents](#) |
- [Author index](#) |
- [Subject index](#) |
- [Search](#) |
- [Home](#)

Texty, a visualization tool to aid selection of texts from search outputs

[Jaume Nualart](#)

Faculty of Arts and Design, Univerity of Canberra & Machine Learning Research Group,
NICTA. Australia

[Mario Pérez-Montoro](#)

Department of Information Science, Faculty of Information Science. University of
Barcelona, Spain

Abstract

Introduction. The presentation of the results page in a search system plays an important role in satisfying the information needs of a user. The usual performance management criteria and tools to organise results have limitations that may hinder the satisfaction of those needs. We present Texty as a new approach that can help improve the search experience of users.

Method. The corpus of texts to which we applied Texty were papers from *Information Research*. To filter the texts we have build five groups of words or vocabularies on concrete fields of knowledge: conceptual approach, experimental approach, qualitative methodology, quantitative methodology and computers/IT.

Results. We show how Texty, intrinsically, is capable of encoding or offer its users information about the text that other alternative classic representations (bar or lines charts, mainly) are not able to offer.

Conclusions. Texty is a complementary tool that improves intellectual interaction with a lists of texts, allowing users to choose texts more effectively knowing their structure before reading them.

CHANGE FONT

Introduction

Information retrieval is a critical factor in an environment characterised by excess of information ([Baeza-Yates & Ribeiro-Neto 2011](#)). When a user conducts a search, the information retrieval systems normally respond with a list of results. In many cases, the presentation of those results play an important role in satisfying user information needs. A bad or inadequate presentation can hinder the satisfaction of the information needs ([Shneiderman 1992](#), [Baeza-Yates 2011](#), [Hearts 2009](#), [Baeza-Yates et al. 2011](#)).

Typically, information retrieval systems present the results of a query in flat, one dimension lists. Usually, these lists are opaque in terms of order, i.e., the users do not know why the list has a particular order. To refine their search, the users have to interact again, normally by filtering the first output of results.

The four main criteria used to organize a list of results are order, relevance, recommendation and clustering ([Morville & Rosenfeld 2006](#), [Pérez-Montoro 2010](#)). The order organises the list of results by alphabetical or numerical order of some of the features (name of the author, date of creation) of the retrieved document. The relevance ranks the retrieved documents considering the relevance of the content of the document to the user's query. The recommendation can sort the results by using the number of recommendations suggested by other users who have previously used this result. The clustering presents the results grouped into a number of subsets formed by documents that deal with the same topic and/or addressing the topic with a similar approach ([Larson 1991](#), [Tryon 1939](#)).

All these forms of organizing results, although used by most systems of information retrieval, have important limitations. The list of results organized by alphabetical or numerical order provides no extra information to help the users decide which of the listed documents can adequately meet their needs. When organising by relevance, the system places documents that could satisfy the information needs of the user at the top, but no extra information on the approach or on the internal structure of the document is provided. In the case of an organisation on the basis of recommendation, the top of the list provides documents recommended by other users and it does not provides extra information on the approach or the internal structure of the document. Finally, clustering provides extra information about the topic of the retrieved document, but it does not guide the user on the distribution and thematic structure of the document.

Visual presentation of search results

In recent years, apart from these more standard presentation of results and with the aim to overcome some of its limitations, a number of visual proposals have been developed to improve user interaction with search results. Most of these proposals can be articulated in three main groups: the clustering visualizations, the visualization of query terms and the visualizations using thumbnail images or miniaturized images of documents.

Visualization of clusterings intends to represent the categories and relations between those categories of the retrieved documents. The main trends in these representations are based on the use of, among others, treemaps, tag clouds or network graphs.

The treemaps represent hierarchical relationships of a set of categories using nested rectangles and optimizing the space used for the visualization ([Shneiderman 1992](#), [Shneiderman 2009](#)). Each rectangle's size is proportional to the number of retrieved documents under that category. Normally the rectangles are coloured according to the category they belong to, for easy reading by users (see Figure 1).

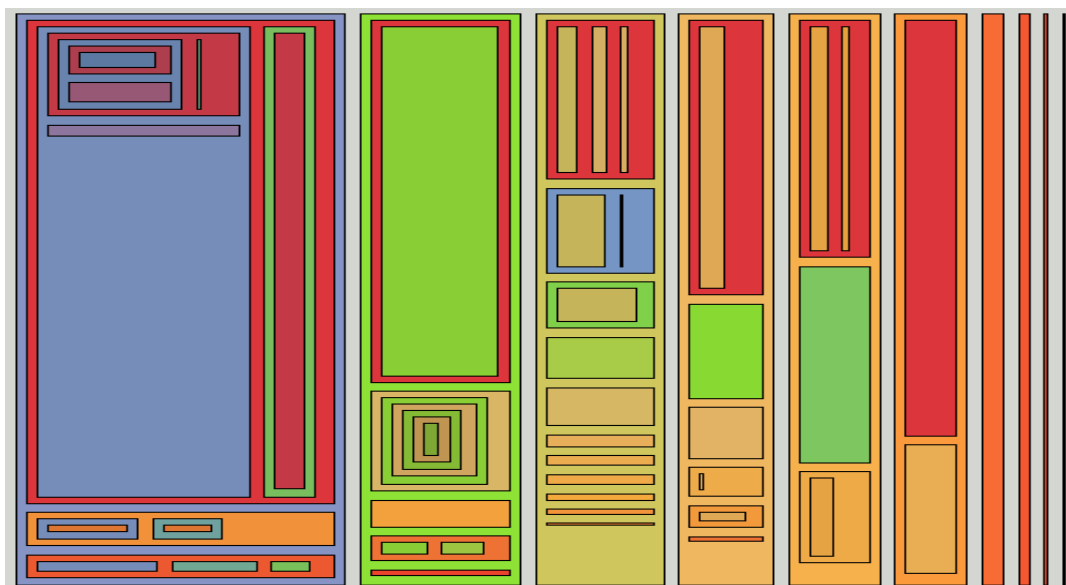
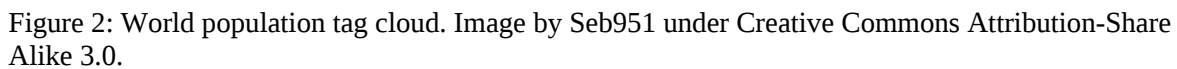


Figure 1: Treemap of a file system.

The tag clouds represent the categories in a group of words, where the color and size of each word are proportional to the number of documents retrieved for each category ([Begelman et al. 2006](#)). The labels that appear in the cloud are usually hyperlinks that lead to the list of documents that have been retrieved under that label.



The visualization of query terms given by the user tend to follow two strategies: visualization of the terms within the document or in a page of results ([Hearst 2009](#)). In the first case, the system outputs the document highlighting those words that literally match the terms of the query ([Egan, et al. 1989](#)). Some studies indicate that users prefer to see this technique implemented by using colour in highlighted words that match the query terms ([Hornbæk and Frøkjær 2001](#)). In the second case, each document is represented on the results page as a horizontal bar proportional to extension of the document and small squares are added for each query terms that appear in the text ([Hoeber and Yang 2006](#)). As in the previous case, some studies indicate that this representation improves with the introduction of a colour scale proportional to the frequency of the query terms in the document ([Anderson et al. 2002](#)).

All these new proposals can improve the search experience of users, but they all have important limitations.

Visualizations based on the query terms also have important limitations. On the one hand, they only provide documents in which the query terms appear. They do not provide extra information on the thematic focus of the retrieved documents, nor possible semantic relationships between retrieved documents. They do not give any orientation on the distribution and structure of those terms unrelated to each of these retrieved document either.

Finally, the visualization strategy which involves completing the list with thumbnail images or miniaturized images of retrieved documents also has important limitations. These visualizations, though complementary, do not provide extra information on the thematic focus of the content of the retrieved documents or on semantic relationships between retrieved documents. They do not show to the user the distribution and thematic structure of each document either. Along these lines, studies show that the thumbnails images strategy does not significantly improve the search experience of users (Czerwinski *et al.* 1999, Dziadosz and Chandrasekar 2002), although they can be helpful in part if the images are enlarged (Kaasten *et al.* 2002)

These limitations lead us to seek new forms of visualization that can help to improve the search experience of users in information retrieval systems and any other case where the user has to choose or select documents from one dimension lists of documents.

The proposed tool presented in this paper aims to face these limitations when deployed as a complement to traditional one dimension list of documents or to a list of results of information retrieval systems, such as clustering or sorting by relevance. This tool shows the essential parts of the contents of each item on the retrieved list and it helps the users in identifying the structure of the content of text documents, without having to tackle each one of the results intellectually.



Figure 3: Use of texty as a complement in a list of results of a search query

We used graphic techniques that were very similar for those used for authorship recognition by Keim and Oelke (2007) and for those used by Hearst (1995) in TileBars. Keim's technique represents the length of the phrases in each text as little squares with colour grading. In this paper this technique is applied very differently. TileBars also show the distribution of terms along the text as Texty does, but the terms come from a search query and the colour intensity is proportional to the frequency of the queried term in each document. In Texty, colour dots show the density of concepts referring to a particular linguistic field, which we call in this paper vocabulary. In Texty,

the human eye analyses visualizations as it would do in Keim and Hearst. Visual coincident factors are colour zones, density of dots and the position and distribution of dots on the plane.

A third tool that graphically is similar to our work is Table Lens ([Rao and Card 1994](#)). In our case, we are not representing table structured data (columns, rows, data in each cell), like Table Lens does. Also Texty is not interactive as Table Lens is. Texty is simpler tool and it does not allow accurate data browsing neither.

Technically, a Texty is an image, an icon that represents the physical distribution of keywords of a text as a flat image. These keywords are grouped in vocabularies, to each of which a colour is linked (see Figure 4). Texty reveals, the structure, conceptual density and subject matter of a text. Texty is a non-intrusive technique, in that an eventual implementation it does not necessarily interfere with the original information system that stores the documents. In this paper we show that this text representation tool enriches the one-dimensional lists that result from searches or from any other static list of documents.



Figure 4: Texty: the process and the colour's legend

It is important to note that the human brain is capable of detecting variations of dots' density (Burguess and Barlow 1983) independently of the used colours ([Nelson and Halberg 1979](#)). Each array of dots of colour may represent a concrete linguistic field or vocabulary. The human vision can differentiate between colours quite well, especially when green and red are not present at once ([Few 2008](#)).

Method

In 2008 and 2009 at the Ludwig Boltzman Institute, Linz, Austria, the challenge was laid down to make visualization tools with the data from the archive of *Ars Electronica*, a file of digital culture, media art and

technology that had been collecting data since 1987. A lack of representation of collections of texts with the same linguistic register was identified. The research was to find the way of representing a text before reading it: a way to distinguish texts on a list and be able to compare them. Initially, as well as texts, there were five high-quality vocabularies on the history of media art: art work, person or institution, date, keyword and award. These were worked out by G. Dirmoser ([Offenhuber and Dirmoser 2009](#)) who provided the basis for developing a tool that showed these five vocabularies by five different colours in a proportional, representative image of the actual text. This work gave the first intuitive representations with the Texty technique.

In this paper we present a more elaborate study of this technique, analysing its features in comparison with classic techniques of representation. The aim of the study was to develop and improve this technique as a complement to information search and retrieval systems.

The stages of the research were: data selection, choice of semantic categories, selecting and identifying the sources for the vocabulary corpus, the processing of terms for each vocabulary, the design of the corpus of texts for representation and the creation of Textys, for each text of the collection.

Data selection

For the study we looked for a controlled collection of texts with a similar register and a specific semantic field. In addition, to assist the study, the texts had to be freely accessible.

For all these reasons we chose the papers published in *Information Research*. These papers belong to the same document collection, have unity, share the academic register, have a similar structure (introduction, method, analysis, results) and have standardized quality (peer-reviewed).

The *Information Research* Website has a search system, by theme, by number or by author. It has a separate list of reviews, along with two retrieval systems: Atomzsite search and Google. The aim of this paper is to present Texty, a tool for information retrieval representation that goes further than the above resources for locating information.

Choice of semantic categories

Once the corpus of texts had been chosen, we identified the following subject categories that could help to classify their contents: conceptual approach, experimental approach, qualitative methodology, quantitative methodology and computers and information technology.

We could have chosen other alternative categories, however, from our personal perspective, as researchers working mainly in IT related issues, the five categories we have chosen are the main criteria for selecting the literature we use to perform the state of the art of the discipline: approach, methodology and degree of technology employed.

The election of the semantic categories, though related to the corpus of texts, is not unique and could be different without affecting the presentation of Texty as a possible helpful tool.

Sources for the corpora of the vocabularies

The next step was to identify the sources of information from which the vocabularies that would be developed subsequently could be extracted. The choice of these sources was based on two complementary criteria. One was the intellectual prestige of the source. This criterion led us to select the [Stanford Encyclopedia of Philosophy](#) and the [Encyclopaedia Britannica](#). A second criterion was the popularity of the source, which led us to choose [Wikipedia](#). The distribution of sources by subject matter is given in Table 1.

Table 1: Concepts and sources of the concepts or the five vocabularies chosen

Definitions				
Qualitative Methodology	Conceptual Approach	Computers & information technology	Quantitative Methodology	Experimental Approach

Stanford Encyclopedia of philosophy	Aristotle's categories	Concepts		
	Intrinsic vs. extrinsic	Category	Mathematics	Experiment in physics
	Properties	Theory	Statistics	
Britanica	Qualitative states			
	qualitative tests to distinguish alternative theories		Mathematics statistics	
Wikipedia	Qualitative data		List of programming languages	
	Quantitative property	Terminology	List of popular computers	Test method
	Qualitative properties	Theory		Case study
	Qualitative research	Vocabulary	List of hardware	Experiment
	Quality (philosophy)	Concept	componets, software glossary	

Importantly, although this has not been implemented in this study, it would be interesting to create, for each of the five vocabularies, a thesaurus (controlled vocabulary) which would spell out the different types of terms (preferred terms, variant terms, broader terms, narrow terms and related terms) and semantic dependencies (equivalence, hierarchy and association) between terms. This solution would solve the problems of silence and noise in indexing derivatives of synonymy and polysemy of terms.

Processing of the terms for each vocabulary

Then the five vocabularies, based on the five corpora of texts for the concepts chosen, were defined (see Table 1). First, a stopwords filter was used, to take out the empty words. Then the words occurring fewer than four times were deleted, as they were considered of little significance for each subject. Then the words occurring in more than one vocabulary were deleted, i.e., we removed interference between vocabularies. Thus we obtained a number of terms for each vocabulary (Table 2).

Table 2: Words selected for each vocabulary before the intellectual review.

Vocabulary	Number of terms	Terms /1,000 words
Conceptual approach	610	21.16
Experimental approach	510	23.29
Qualitative methodology	451	19.71
Quantitative methodology	700	22.24
Computers & information technology	312	18.91

Finally, there was an intellectual review to detect terms that were inconsistent with the subject matter, ambiguous terms and terms that were not coherent with each vocabulary. The experimental vocabularies were configured as in Table 3.

It should be said that the objective of this paper was to introduce the potential of the Texty tool. It is not a goal of this paper to study the best strategies to define the words that best represent a concrete field of knowledge or, as we call it in this paper, a vocabulary. In machine learning there are very powerfull methods like building topic models. This is an interesting possibility for future research.

Table 3: Final number of terms for each vocabulary.

Vocabulary	Number of terms
Conceptual approach	65
Experimental approach	53
Qualitative methodology	74
Quantitative methodology	86
Computers & information	110

The computers and information technology vocabulary is descriptive, which is why we left a large number of terms, as they all clearly refer to computers and information technology. The final list of words in all the vocabularies [can be found here](#).

Regarding the possible overlapping of colour dots we recommend to set a number of terms per vocabulary that avoids an excess of terms per line of the text.

Corpora of texts to represent

After choosing the papers from *Information Research* as corpora of texts to which the Texty tool can be applied, then a private replica of the journal *Information Research* was made to conduct the study in laboratory comfort and also to show clearly how Texty can be implemented in an existing system. *Information Research* is made out of static HTML pages and Texty has been introduced in each issue's index and in subject index (link in red on the left column). Texty representations of papers in *Information Research* papers [are to be found here](#). [To access, use the user name 'texty' and the password 'texty'].

Creation of the Textys

There are a lot of ways to, technically, create Textys. The choice it will depend on the required level of production and concrete conditions of each case. Here we describe the simple automated method that has been used to create the almost 500 Textys required for this study.

The initial format of the texts taken from the *Information Research* website was HTML. The HTML files were parsed and an specific class for each vocabulary was applied to all vocabulary's words found. Then a specific colour's style was defined using a Cascade Style Sheet (CSS), to the vocabulary's words; the rest of text was defined as white. Finally, a screenshot of the HTML page was taken and the size was adjusted to 300x450px for each Texty.

In this study we created Textys with five different colours, as shown in Figure 5. When choosing the colours, the main restrictions usually recommended for this kind of graphic attribute were taken into account. One restriction was the use of basic colours that most of humans can distinguish ([Kay and Maffi 2008](#)). A second restriction was that humans have very little difficulty identifying three to five colours; and for seven to nine colours the identification becomes significantly more difficult ([Healey 1996](#)).

Legend

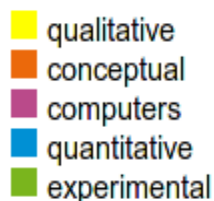


Figure 5: Colours of the vocabularies of Texty

At this point we should stop a moment to analyse the information contained in the white areas of a Texty. Since the Texty is a physical representation of data, i.e., the colour dots appear in positions that reflect the real positions of terms in the text, the absence of ink gives relevant information about the text represented. Bearing in mind the theory of Tufte on the ratio of ink and data ([Tufte 1986](#): 93), for Texty we would have to say that they are data without ink. The absence of colours means a lower density of terms of the proposed vocabularies along the text. If we view the white zones as zones with data, Tufte's formula in the case of Texty would be as follows (see Figure 6), with the maximum proportion of ink devoted to representing data:

$\text{Data-ink ratio} = \frac{\text{Data-ink}}{\text{Total ink used to print the graphic}} =$ <p style="margin-left: 40px;">= Proportion of a graphic's ink devoted to non-redundant display of data-information =</p> <p style="margin-left: 40px;">= 1.0 - Proportion of graphic that can be without loss of data-information</p>	$\text{Data-ink ratio} = \frac{100}{100} = 1$
--	---

Figure 6: Tufte's data-ink ratio equation (left) and the application to Texty (right)

Results

We created Textys for all the papers in the *Information Research*, from Volume 1, No. 1 (1995/96) to Volume 15 No.4 (2010), with a total of 454 Textys. Below, Figure 7 gives the Textys of the 17 papers in [volume 15, no. 4](#) (December 2010) of the journal.


















<p>1. Proceedings of ISiC</p>  <p><i>Cultural differences in the health information environments and practices between Finnish and Japanese university students</i> Graeme Baxter, Rita Marcella and Laura Illingworth</p>	<p>2. Proceedings of ISiC</p>  <p><i>Organizational information behaviour in the public consultation process in Scotland</i> Leanne Bowler</p>	<p>3. Proceedings of ISiC</p>  <p><i>Talk as a metacognitive strategy during the information search process of adolescents</i> Jenny Bronstein</p>	<p>4. Proceedings of ISiC</p>  <p><i>Selecting and using information sources: source preferences and information pathways of Israeli library and information science students of your paper</i> Donald O. Case</p>
<p>5. Proceedings of ISiC</p>  <p><i>A model of the information seeking and decision making of online coin buyers</i> Kreetta Askola, Toshimori Atsushi and Maije-Leena Huotari</p>	<p>6. Proceedings of ISiC</p>  <p><i>Local versus global information relevance in Website use: a case study with the information literacy portal ALIAHIES</i> Francisco Javier García Marco and Maria Pinto</p>	<p>7. Proceedings of ISiC</p>  <p><i>Information behaviour research and information systems development: the SHAMAN project, an example of collaboration</i> Elena Maceviciute and T.D. Wilson</p>	<p>8. Proceedings of ISiC</p>  <p><i>Avoiding health information in the context of uncertainty management</i> Anu Sairanen and Reijo Savolainen</p>
<p>9. Proceedings of ISiC</p>  <p><i>A study of labour market information needs through employers' seeking behaviour</i> Sonia Sanchez-Cuadrado, Jorge Morato and Yorgos Andreidakis</p>	<p>10. Proceedings of ISiC</p>  <p><i>Information in context: co-designing workplace structures and systems for organizational learning</i> Mary M. Somerville and Zaena Howard</p>	<p>11. Proceedings of ISiC</p>  <p><i>"We have a lot of information to share with each other". Understanding the value of peer-based health information exchange</i> Tiffany C. Veinot</p>	<p>12. Proceedings of ISiC</p>  <p><i>Information sharing: an exploration of the literature and some propositions</i> T.D. Wilson</p>
<p>13. Proceedings of ISiC</p>  <p><i>Applying McKenzie's model of information practices in everyday life information seeking in the context of the menopause transition</i> Alison Yeoman</p>	<p>14. Regular paper</p>  <p><i>Double or nothing: is redundancy of spatial data a burden or a need in the public sector of Uganda?</i> Walter T de Vries and Beatrice Winnie Nyemera</p>	<p>15. Regular paper</p>  <p><i>Analysis of automatic translation of questions for question answering systems</i> Loia García-Santiago and Maria-Dolores Olivera-Lobo</p>	<p>16. Regular paper</p>  <p><i>Dietary blogs as sites of informational and emotional support</i> Reijo Savolainen</p>
<p>17</p>  <p><i>Information and information science: an address on the occasion of receiving the award of Doctor Honoris Causa. at the University of Murcia, 30 September, 2010 T.D. Wilson</i></p>	<p>Legend</p> <ul style="list-style-type: none"> qualitative conceptual computers quantitative experimental 		

Figure 7: Textys for *Information Research*, volume 15, No 4 (December 2010) and legend.

We presented Texty as a simple and complementary tool to enrich lists of texts. In this respect, a first glance at Figure 7 can help the reader to select papers to read as follows:

- The predominant tone in this issue is experimental (green), though followed closely by the qualitative approach (yellow).
- Papers 3, 11 and 13 look clearly experimental (green), while paper 7 looks like one that requires the reader to have more knowledge of computers and information technology (violet).
- Five of the seventeen papers (38.5%) have a notable presence of computers and information technology (violet).
- The paper with the biggest conceptual load is the 9th, though the 7th, 8th and 16th also have a conceptual content (orange).
- The more generalist paper seems to be the 15th.
- This issue does not involve quantitative methodologies much (blue).

Here we can see how Texty can be used for the exploration and navigation of texts before they are read.

Starting from this development we want to see what would happen if we try to represent the same data (terms and vocabularies from papers) using traditional techniques, like bar and lines charts. We are not proposing bar and line charts to be used in the same way as Texty, i.e., to enrich lists of texts, but we are comparing formally how the same data set would look under these techniques compared to Texty technique.

Texty and the bar charts:

To illustrate this comparison, we chose papers from Volume 15, No. 4 (December 2010).

Case one, paper 441: [A study of labour market information needs through employers' seeking behaviour](#). Sonia Sanchez-Cuadrado, Jorge Morato, Yorgos Andreidakis and Jose Antonio Moreira

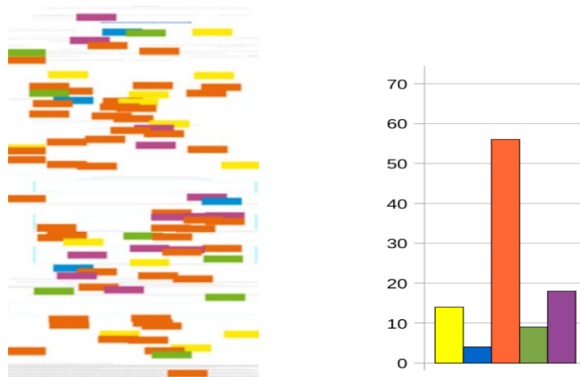


Figure 8: Texty and bar chart for paper 441 (Information Research).

Both methods identify the most common vocabulary. In this case it is the conceptual one (orange). This paper describes knowledge representation techniques with computer support, which the two representations also show us. However, in the case of Texty, it can be seen that these techniques are discussed in the middle part of the paper (violet colour), whereas this was not seen with the bar chart.

Case two, paper 445: [Information behaviour research and information systems development: the SHAMAN project, an example of collaboration](#). Elena Maceviciute and T.D. Wilson

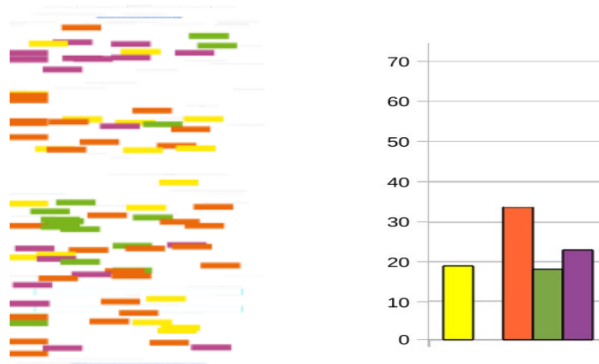


Figure 9: Texty and bar chart for paper 445 (Information Research).

This paper has a conceptual tone (orange). Initially, in the background on long-term digital preservation, we can say that techniques that require computers are being discussed (for example: e-mail, word-processed documents and spreadsheets, as well as e-books, sound recordings, films, scientific data sets, social science data archives, are terms used in this paper). In the middle of the paper we saw a concentration of green points belonging to the experimental vocabulary. This coincides with the explanation of the data used by the SHAMAN program on the basis of interviews with users. Not all this information can be deduced from the bar chart.

Case three, paper 450: [Analysis of automatic translation of questions for question-answering systems](#). Lola García-Santiago and María-Dolores Olvera-Lobo

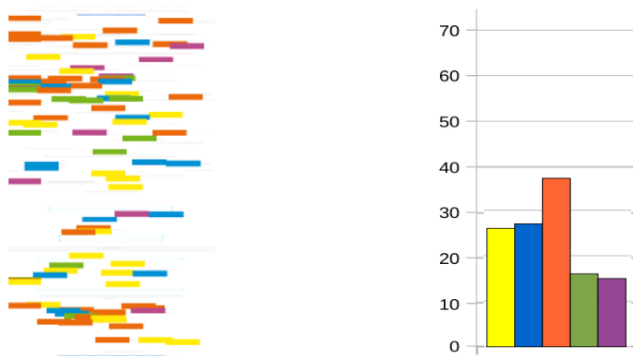


Figure 10: Texty and bar chart for paper 450 (Information Research).

In this case we have a paper with a considerable presence of the five vocabularies. Here, the importance of being able to see the physical distribution of terms in the paper can be seen perhaps in greater clarity. Thus we can say that the paper starts with a conceptual tone, to then explain the method in an experimental tone. The paper does not require too much knowledge of data processing, although there are references to it in the first half. At the end there are references of a conceptual kind. In general, the paper has a qualitative approach, as yellow is distributed throughout. Again, none of this information can be extracted by the bar chart.

Texty and the line charts:

We used a line chart with the following coordinates: Y axis represents the position of the first character of the term. The X axis represents the number of terms for each vocabulary. Figure 11 gives an example of this representation.

Case four, paper 438: [Dietary blogs as sites of informational and emotional support](#), by Reijo Savolainen

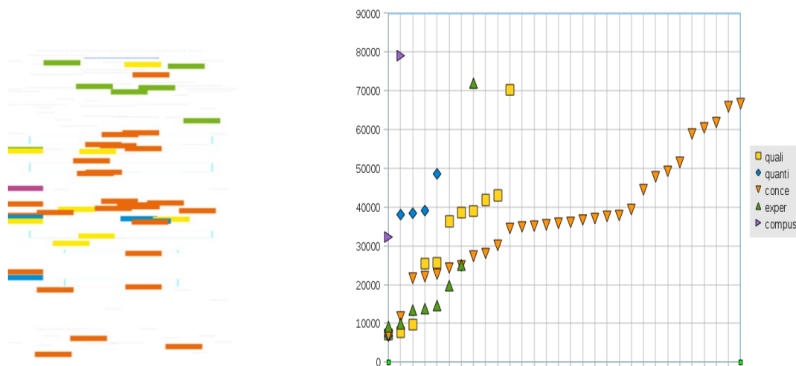


Figure 11: Texty and line charts for paper 438 (Information Research).

The reading of the line chart contributes more information on the structure and distribution of terms in the text than the bar chart does. Line chart shows very well the number of terms of each vocabulary. Even the line chart is more accurate in number of vocabularies than Texty. However, Texty is more suitable for everywhere use because it doesn't require the use of axes and has a bigger range of readable sizes. For small sizes the line chart axes scales becomes unreadable.

Texty versus bar and line charts

The objective of our work is not to quantitatively study Texty's performance against other visualizations, taking this into account, both options, Texty and the charts, show the number of terms in each vocabulary, i.e., the general focus of a paper at a glance. The improvements introduced by Texty are:

1. Texty shows the distribution of terms along the text.
2. With Texty the conceptual structure of the paper can be seen: e.g. at the start there is a conceptual explanation; then the experimental part is developed; finally, the calculations in which there is intensive use of technology related and/or computer related operations.
3. Texty doesn't need axes or coordinates and scales.

Conclusions

The development of this work allows us to point out a number of learned lessons about the complementary nature of Texty, its ability to encode information, and its non intrusive structure and technology.

Because of its complementary nature, Texty enriches lists of texts adding an image that physically represents the distribution of five conceptual fields along the text. Texty is not a replacement for classic search systems, but is proposed as a complement.

Texty's ability of encoding means that it is able to present distribution and structure of a text using only coloured dots that represent the text itself.

Another conclusion we can draw is that Texty is not an intrusive solution from the point of view of the architecture of information. In this sense, Texty can be implemented without affecting the organizational criteria (e.g. order, relevance, recommendation or clustering) used to produce the retrieved list of documents.

Texty is a tool that can be implemented in an existing collection of texts and in is non-intrusive from a technological point of view. That means that it is not necessary to change or reprogram the storage system where the collection of texts lies. This easy implementation is presented as a critical advantage for future Texty implementations.

Finally, strictly speaking, from the point of view of information retrieval, the use of Texty is not adding any advantage (not improved indexing and search algorithms, for example). What Texty aims to do is to improve is the presentation of results: it complements the traditional list of results (generally based on a title and a short summary) providing information on the content and structure of the retrieved document without having to interact directly with the document itself (see Figure 3).

Future developments

We want to round off these conclusions by mentioning some future lines of development derived from Texty.

Texty can be exported to other backgrounds and other vocabularies, adapted to each case and it can be personalized to the extent that it shows us other vocabularies (colours) depending on the reader preferences or the texts represented. Representation can be expanded and texts sections separators added, which indicate, for example, the customary sections of an paper (intro, method, analysis, results, conclusions, in the case of the papers of *Information Research*).

Dynamic, personalized and folk-vocabularies can increase the efficacy of Texty, as can the use of different layers to represent any vocabularies, as wanted. The use of interactive images (sensitive to clicks on the mouse) allows Texty to navigate through the text in question.

As noted, the use of thesaurus would improve the representative capacity of the vocabularies used in texty. The adaptation of Texty for texts in a number of languages is another possible use: all you need are translations of the vocabularies.

Acknowledgements

Our thanks to: Jordi K. Nualart, Amelia S. Nascimento, Mar Canet (Catalonia), Sandor Herramhof (Linz, Austria), Dietmar Offenhuber (MIT USA), Joelle Vandermensbrugghe (University of Canberra), and the Department of Information Science, Faculty of Information Science. University of Barcelona.

About the authors

Jaume Nualart is a PhD candidate in the Faculty of Arts and Design, University of Canberra and a research engineer at NICTA, Australia. MAS and MSc (Licenciatura) at University Autonomus of Barcelona He can be contacted at Jaume.Nualart@canberra.edu.au.

Mario Pérez-Montoro is a Professor in the Department of Information Science, at University of Barcelona, Spain. He completed his Bachelor's degree in Philosophy and Education from the University of Barcelona (Spain), his Master of Information Science in Organizations from the Politechnical University of Catalonia (Spain), and his PhD from the University of Barcelona (Spain). He has been visiting scholar at the Center for the Study of Language and Information (CSLI) at Stanford University (California, USA) and at the School of Information at UC Berkeley (California, USA). He can be contacted at: perez-montoro@ub.edu.

References

- Anderson, T. J., Hussam, A., Plummer, B. & Jacobs, N. (2002). Pie charts for visualizing query term frequency in search results. *Proceedings of the 5th International Conference on Asian Digital Libraries: Digital Libraries: People, Knowledge, and Technology* (pp. 440–451). London: Springer-Verlag.
- Baeza-Yates, R. (2011). [Tendencias en recuperación de información en la web](#). [Trends in information retrieval on the Web.] *BiD: textos universitaris de biblioteconomia i documentació*, desembre, núm. 27. Retrieved from <http://www.ub.edu/bid/27/baeza2.htm> on 22-01-2013.
- Baeza-Yates R. A. & Ribeiro-Neto, B. (2011). *Modern information retrieval*. Boston, MA: Addison-Wesley Longman.
- Baeza-Yates, R., Broder, A. & Maarek, Y. (2011). The new frontier of Web search technology: seven challenges. In S. Ceri & M. Brambilla (Eds.), *Search Computing* (Vol. 6585, pp. 3–9). Berlin & Heidelberg: Springer Verlag.
- Begelman, G., Keller, P., Smadja, F. & others. (2006). *Collaborative Web Tagging Workshop at WWW2006*, Edinburgh, Scotland (pp. 15–33). Retrieved 2 June, 2013 from <http://www.ra.ethz.ch/cdstore/www2006/www.rawsugar.com/www2006/20.pdf> (Archived by WebCite® at <http://www.webcitation.org/6H4xIzT55>)
- Brandes, U., Hofer, M. & Lerner, J. (2006). [WordSpace: visual summary of text corpora](#). In Robert F. Erbacher, Jonathan C. Roberts, Matti T. Gröhn & Katy Börner (Eds.). *Visualization and data analysis 2006* (pp. 212-223). Bellingham, WA: SPIE-the International Society for Optics and Photonics.

- (Proceedings of SPIE, Volume 6060). Retrieved 2 June, 2013 from <http://www.mpi-inf.mpg.de/~mhoefer/05-07/Brandes06WordSpace.pdf>.
- Burgess, A. & Barlow, H. B. (1983). The precision of numerosity discrimination in arrays of random dots. *Vision Research*, 23(8), 811–820.
 - Czerwinski, M., Van Dantzich, M., Robertson, G. & Hoffman, H. (1999). [The contribution of thumbnail image, mouse-over text and spatial location memory to web page retrieval in 3D](#). In *Proceedings of the INTERACT'99 conference*, (pp. 163-170). Dordrecht, Kluwer. Retrieved 2 June, 2013 from <http://research.microsoft.com/en-us/um/people/marycz/interact99.pdf>
 - Dziadosz, S. & Chandrasekar, R. (2002, August). Do thumbnail previews help users make better relevance decisions about web search results?. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, (pp. 365-366). New York, NY: ACM Press.
 - Egan, D.E., Remde, J. R., Gomez, L.M., Landauer, T.K., Eberhardt, J. & Lochbaum, C.C. (1989). Formative design evaluation of superbook. *ACM Transactions on Information Systems (TOIS)*, 7(1), 30–57.
 - Fellbaum, C. (2010). WordNet. In Roberto Poli, Michael Healy & Achilles Kameas, (Eds.). *Theory and applications of ontology: computer applications*, (pp. 231-243). Berlin & Heidelberg: Springer
 - Few, S. (2008, February). [Practical rules for using color in charts](#). *Visual Business Intelligence Newsletter*, No. 11. Retrieved 2 June, 2013 from <http://www.perceptualedge.com/library.php>
 - Granitzer, M., Kienreich, W., Sabol, V., Andrews, K. & Klieber, W. (2004). Evaluating a system for interactive exploration of large, hierarchically structured document repositories. In *IEEE Symposium on Information Visualization, 2004. INFOVIS 2004*. on (pp. 127-134). Los Alamitos, CA: IEEE Computer Society Press.
 - Healey, C. G. (1996). Choosing effective colours for data visualization. In *Proceedings of the 7th conference on Visualization '96* (p. 263–ff.). Los Alamitos, CA, USA: IEEE Computer Society Press.
 - Hearst, M. (1995). TileBars: visualization of term distribution information in full text information access. In *CHI '95: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (pp. 59-66). New York, NY: ACM Press/Addison-Wesley Publishing Co.
 - Hearts, M. (2009). *Search user interfaces*. Cambridge: Cambridge University Press.
 - Hoeber, O. & Yang, X. D. (2006). A comparative user study of web search interfaces: HotMap, Concept Highlighter, and Google. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, (pp. 866-874). Washington, DC: IEEE Computer Society.
 - Hornbæk, K. & Frøkjær, E. (2001). Reading of electronic documents: the usability of linear, fisheye, and overview+ detail interfaces. In *CHI '01 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (pp. 293-300). New York, NY: ACM Press.
 - Jhaveri, N. & Räihä, K. J. (2005). The advantages of a cross-session web workspace. In *CHI EA '05 CHI '05 Extended Abstracts on Human Factors in Computing Systems*, (pp. 1949-1952). New York, NY: ACM Press.
 - Kaasten, S., Greenberg, S. & Edwards, C. (2002). How people recognise previously seen Web pages from titles, URLs and thumbnails. In Kristine Faulkner, Janet Finlay & Françoise Détienne *People and Computers XVI - Memorable Yet Invisible: Proceedings of HCI 2002* (pp. 247–266). Berlin/Heidelberg: Springer.
 - Kay, P. & Maffi, L., (2008). Number of basic colour categories. In Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie, (Eds.). *The world atlas of language structures online*. (ch. 133). Munich, Germany: Max Planck Digital Library.
 - Keim D. A. & Oelke D. (2007). Literature fingerprinting: a new method for visual literary analysis. In *VAST '07: Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology*, (pp. 115-122). Washington, DC: IEEE Computer Society
 - Larson, R. R. (1991). Classification clustering, probabilistic information retrieval, and the online catalog. *The Library Quarterly*, 61(2), 133–173.
 - Moya-Anegón, F., Vargas-Quesada, B., Herrero-Solana, V., Chinchilla-Rodríguez, Z., Corera-Álvarez, E. & Muñoz-Fernández, F. J. (2004). A new technique for building maps of large scientific domains based on the cocitation of classes and categories. *Scientometrics*, 61(1), 129–145.
 - Morville, P. & Rosenfeld, L. (2006). *Information architecture for the world wide web: designing large-scale web sites*. Sebastopol, CA: O'Reilly Media.
 - Nelson, M. & Halberg, R. (1979). Visual contrast sensitivity functions obtained with colored and achromatic gratings. *Journal of the Human Factors and Ergonomics Society*, 21(2), 225-228.
 - Offenhuber D. & Dirmoser G. (2009) [Semaspaces: graph editor for large knowledge networks](#). Retrieved 19 January, 2013 from <http://residence.aec.at/didi/FLweb/>. (Archived by WebCite® at <http://www.webcitation.org/6DmoMu1n3>)

- Pérez-Montoro, M. (2010). *Arquitectura de la información en entornos web*. El profesional de la información, **19**(4), 333-338.
- Rao, R. & Card, S. (1994). The table lens: merging graphical and symbolic representations in an interactive focus+ context visualization for tabular information. In *CHI '94 Conference Companion on Human Factors in Computing Systems*, (pp. 222). New York, NY: ACM Press.
- Shneiderman, B. (1992). *Designing the user interface: strategies for effective human-computer interaction*. (2nd ed.) Boston, MA: Addison-Wesley Longman Publishing Co., Inc.
- Shneiderman, B. (1992). Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics*, **11**(1), 92–99.
- Shneiderman, B. & Plaisant, C. (2009). [Treemaps for space-constrained visualization of hierarchies](http://www.cs.umd.edu/hcil/treemap-history/). Retrieved 3 June, 2013 from <http://www.cs.umd.edu/hcil/treemap-history/> (Archived by WebCite® at <http://www.webcitation.org/6H6Mp735I>)
- [Texty representations of Information Research](#) papers (2011). Restricted access: user:'texty', password: 'texty'.
- Tryon, R. (1939). *Cluster analysis*. New York, NY: McGraw-Hill
- Tufte E. R. . 1986. *The visual display of quantitative information* (pp 93). Cheshire, CT: Graphics Press
- Vié gas, F. B., Wattenberg, M., Van Ham, F., Kriss, J. & McKeon, M. (2007). Manyeyes: a site for visualization at internet scale. *Visualization and Computer Graphics, IEEE Transactions*, **13**(6), 1121–1128.
- Woodruff, A., Faulring, A., Rosenholtz, R., Morrisson, J. & Pirolli, P. (2001). Using thumbnails to search the Web. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 198–205). New York, NY: ACM Press.
- Yee, K. P., Fisher, D., Dhamija, R. & Hearst, M. (2001). Animated exploration of dynamic graphs with radial layout. In *INFOVIS '01 Proceedings of the IEEE Symposium on Information Visualization 2001*, (p. 43). Washington, DC: IEEE Computer Society.

How to cite this paper

Nualart, J. Pérez-Montoro, M (2013). Texty, a visualization tool to aid selection of texts from search outputs. *Information Research*, **18**(2) paper 581. [Available at <http://InformationR.net/ir/18-2/paper581.html>]

Find other papers on this subject

Scholar Search

Google Search

Bing

Check for citations, [using Google Scholar](#)

Like { 0 } Tweet { 0 }

14

754

© the authors, 2013.

Last updated: 3 June, 2013

- [Contents](#) |
- [Author index](#) |
- [Subject index](#) |
- [Search](#) |
- [Home](#)

A.3. Paper III

Citation:

Perez-Montoro, Mario; Nualart, Jaume (2015). Visual articulation of navigation and search systems for digital libraries, *International Journal of Information Management*, Volume 35, Issue 5, October 2015, Pages 572-579, ISSN 0268-4012

Afiliation (in catalan)

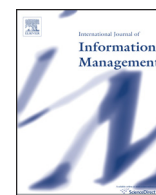
- 1r autor: Mario Pérez-Montoro, doctor i professor al Dept. de Ciències de la Informació de la Universitat de Barcelona.
- 2n autor: Jaume Nualart Vilaplana es doctorand a la Facultat de Biblioteconomia i Documentació de la Universitat de Barcelona, inginyer de recerca al NICTA (Australia), i doctorand a la Faculty of Arts and Design, University of Canberra (Australia).

Summary (in Catalan)

Els portals de revistes científiques i biblioteques digitals són els sistemes d'informació als quals els investigadors recorren més sovint per a la consulta i difusió de llur treball acadèmic. No obstant això, no se n'han millorat les interfícies. Proposem articular els sistemes de navegació i recerca en una sola solució visual que permeti simultàniament l'exploració i la cerca d'un sistema d'informació. Àrea és una eina de visualització de baix cost que és fàcil d'implementar, i que es pot utilitzar amb grans col·leccions de documents. D'altra banda, té una corba d'aprenentatge curta que millora tant l'experiència com la satisfacció dels usuaris quan usen llocs web de revistes científiques i biblioteques digitals.

Access online (URL):

- <http://dx.doi.org/10.1016/j.ijinfomgt.2015.06.005>
- <http://www.sciencedirect.com/science/article/pii/S0268401215000614>



Visual articulation of navigation and search systems for digital libraries



Mario Pérez-Montoro^a, Jaume Nualart^{a,b,c,*}

^a Department of Information Science, University of Barcelona (Spain), Melcior de Palau 140, 08014 Barcelona, Spain

^b University of Canberra Bldg, Floor & Room: 9, C12, Faculty of Arts and Design, University of Canberra, Bruce 2601 ACT, Australia

^c Machine Learning Research Group, National ICT Australia (NICTA) Canberra, Australia

ARTICLE INFO

Article history:

Received 17 May 2015

Received in revised form 22 June 2015

Accepted 24 June 2015

Keywords:

Searching

Browsing

Information visualization

Information management

Digital libraries

ABSTRACT

Journal and digital library portals are the information systems that researchers turn to most frequently for undertaking and disseminating their academic work. However, their interfaces have not been improved. We propose an articulation of the navigation and search systems in a single visual solution that would allow the simultaneous exploration and interrogation of the information system. *Area* is a low-cost visualization tool that is easy to implement, and which can be used with large collections of documents. Moreover, it has a short learning curve that enhances both user-experience and user-satisfaction with journal and digital library websites.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

When designing a digital information system, the first objective that has to be met is that of facilitating the most intuitive means for users of locating information. To satisfy this objective, the systems of organization, labeling, navigation and searching have to be properly designed, as do the controlled vocabularies that articulate this digital environment (Morville and Rosenfeld, 2007).

For a web page, for example, this means that the organizational systems must serve to structure and organize website content. They are usually constructed by using a classification, based on one or more specific criteria of the content housed on that page (for example, the subject that is being dealt with, the date of creation or the audience being targeted). The labeling system consistently and efficiently defines and determines the terms used to name the categories, options and links used on the web in a user-friendly language. The navigation system allows users to move comfortably around the different sections that make up the website. It provides a method of orientation for users to move in a controlled way from one point of the website to another and to ensure that at all times they know where they are and where they can go within the structure of the web. Based on a previous indexing strategy, the

search system allows the user to formulate queries and to retrieve information from within the website. Controlled vocabularies or languages are documentary resources (thesauri, taxonomies, synonym rings, etc.) that facilitate, by articulating the other elements of the architectural structure, the search and retrieval of information on the site (Pérez-Montoro, 2010).

While all these elements form part of the architectural anatomy of a digital information system, the two elements used most frequently by users when seeking information are the search and navigation systems. These two systems tend to be clearly identified in the system interface using the search box and the navigation bar, respectively. Users are typically well versed in their use and, to improve their performance, they are usually articulated via the labeling system (i.e., the navigation system labels are used as indexing terms in the search engine).

In the case of journals and digital libraries, in common with other digital information systems, architectural elements are usually employed to facilitate user location of the information they manage.

Among these elements, the most frequently used are typically their navigation and search systems. In this case, the navigation system is usually quite simple, allowing an exploration of the resources filtered through such criteria as author, year of publication, journal or publisher and, in the best of cases, subject. The results of this navigation appear as a list of clickable labels that lead the user to the set of resources, listed alphabetically, corresponding to these criteria. Search systems usually allow the formulation of queries (e.g.,

* Corresponding author. Fax: +61 262 015 300.

E-mail addresses: perez-montoro@ub.edu (M. Pérez-Montoro), jaume.nualart@canberra.edu.au (J. Nualart).

any word, all words or exact phrase) by field (e.g., title, description, keywords or anywhere). The result of the query is a list of resources; normally sorted alphabetically too; which corresponds to the criteria in the search interface.

These architectural systems and their interfaces are typically adapted to the nature of the documents managed by these systems and to the metadata used. The documents are static, non-dynamic, resources as far as their content is concerned, and they do not change over time. Moreover, their metadata describe the contents stored (based on qualitative, ordinal, nominal or hierarchical data) (Hearst, 2009; van Hoek and Mayr, 2014).

These systems are the direct heirs of the classical interfaces of the document databases on CD-ROM developed in the eighties and which have barely evolved since. In contrast with other information systems, such as e-commerce websites, their interfaces have not been improved on the basis of the findings provided by user studies, nor have the advances developed in specific disciplines, such as information architecture, or those derived more generally from user experience (UX), been applied to them.

2. Visualization of information in digital libraries

One of the options for improving classical interfaces is the introduction of new visual solutions in the search process that improve user-experience and user-satisfaction with these digital systems of scientific information.

Traditionally, following on from the initial query, the search systems implemented in information systems of this type offer a very simple representation of the results retrieved. They usually only provide a vertical list of results sorted alphabetically, and, for each result, they give additional information about the retrieved item, such as its author, the title or date of publication of the document, among others.

This strategy of traditional representation has significant limitations. On the one hand, it does not always provide sufficient information about the content of the document to enable the user to accept it or dismiss it without having to read or interact with it first (Baeza-Yates, 2011; Nualart et al., 2014). And, on the other, it does not allow the user to deploy techniques of berrypicking in the search process (Bates, 1989), which could refine the results obtained so as to propose subsequent, more efficient searches in keeping with the user's changing information needs following interaction with the results.

In an attempt at overcoming these limitations, from the late eighties onward, a series of prototypes have been developed that seek to improve the visualization of results from journal and digital library portals. Some have focused on the representation of the content of the retrieved documents (Hearst, 1995; Egan et al., 1989; Weiss-Lijn et al., 2001; Woodruff et al., 2001; Lam and Baudisch, 2005; Hoeber and Yang, 2006; Nualart and Pérez-Montoro, 2013); while others have contributed new interactive visualizations of the set of results after formulating the search query.

If we focus on the second group of prototypes, we can identify two main types of strategy, some of which are interactive: first, those that provide support for query creation and refinement and, second, those that offer visual support for the presentation of results.

The earliest techniques were designed to help the user in formulating the query, facilitating the use of Boolean operators (Jones, 1998; Wong et al., 2011) or supplying and suggesting possible terms to the user for building their queries (Schatz et al., 1996).

Those focusing on the visual presentation of results include different alternatives. Some offer two-dimensional visualizations of the relationships between the retrieved documents by using maps or clusters (Chalmers et al., 1992; Andrews et al., 2001,

2002) or by using two-dimensional tables or grids (Fox et al., 1993; Shneiderman et al., 2000; Kim et al., 2011). Others present strategies based on three-dimensional visualizations of the retrieved results (Robertson et al., 1991; Hearst and Karadi, 1997; Cugini et al., 2000). These visual prototypes made a series of significant improvements to the classical interfaces of journal and digital library portals. Thus, on the one hand, they provided more rapid search times compared to those of traditional non-visual methods (Hienert et al., 2012) and, on the other, they permitted a more efficient formulation of queries in a way that was tailored to the information needs of users. And, finally, they provided additional information to users, information that was not available on a page of more conventional results. This extra information, which shows different semantic relationships between the documents retrieved, provides a better interaction with the results and facilitates the refinement of subsequent queries (Bauer, 2014).

Yet, even with these advantages, these prototypes and advances in visualization have not been widely implemented in the portals or websites of journals or digital libraries. The reasons for this are varied, but they can be classified into two main groups: reasons of a practical nature and methodological reasons.

In the case of the practical reasons, in resources of this type these tools are implemented as separate pages from the basic search interfaces, which means users perceive them as being secondary tools. Furthermore, these solutions, especially those that visualize the results, involve a high level of abstraction and conceptualization that means they are not very intuitive for users. And, perhaps more importantly, implementing these techniques, unlike traditional interfaces, does not offer any clear commercial or economic benefits in the world of digital systems of scientific information of this type.

If we focus on the methodological reasons, it can be seen that very few of the proposed techniques have been tested and evaluated with end users, which makes it difficult to draw any clear conclusions about their efficiency. Moreover, the prototypes have only been used with small collections of documents, and so their efficient use with large collections has not been demonstrated to users. Likewise, the paucity of the quantitative results reported in these studies of visual prototypes fails to demonstrate whether they are any better than the classical versions of the interfaces. As such, experiments are needed that analyze a period of widespread use over a broader period of time before it can be concluded whether or not the difficulty in using them stems from the users' learning curve and their degree of familiarity with the system. Similarly, when these prototypes are constructed by articulating different techniques it becomes more difficult to compare them, because it is not possible to attribute unequivocally the success or failure of the system to one or more of the techniques implemented. And, in this sense, these tools do not share a methodological design that would allow us to compare the results of each proposal and to analyze them jointly.

3. Area: an alternative visualization proposal

To overcome these practical and methodological limitations, new solutions and low-cost tools that can be readily implemented, and which can improve user-experience and user-satisfaction with these information systems, need to be identified. One possible alternative is the articulation of the navigation and search systems in a single visual solution that would allow the simultaneous exploration and interrogation of the information system.

Area is a new, low-cost visualization tool that is easy to implement, and which can be used with large collections of documents. Moreover, it has a short learning curve that articulates the two systems using a two-dimensional structure that can enhance both

Table 1
Comparison of features of existing Information Research site and *Area*.

Features	Information Research	Area	Comment
Explore by issue as a list of papers	Yes	Yes	No changes: Area redirects to the existing issue page
Search with Atomz, and search with Google	Yes	Yes	No changes: it redirects to the IR search page
Multiples overviews of the collection	No	Yes	New feature: (no. of eligible properties) 2 This is 52 = 25 combinations of eligible properties
Numerical overview	No	Yes	New feature: Area shows an overview of the main numbers of the collection
Topic distribution	No	Yes	New feature: filter papers are marked during exploration.
Explore by language	No	Yes	New feature: language is an eligible property. So it can be represented in combination with the other of properties.
Explore by year, issue and volume	Yes	Improved	Improved feature: multiple representations and evolution visualization
Explore by subject	Yes	Improved	Improved feature: TAB “by topic”, allows filter by typing
Explore by author (authors can have more than one paper)	Yes	Improved	Improved feature: TAB “by author”, allows filter by typing
How many papers talk about a subject?	Yes	Improved	Improved: Area shows the papers and their context. A better visualization of the group of results

user-experience and user-satisfaction with journal and digital library websites.

Although the idea for *Area* originated in 2006, it has evolved since then with the development of versions in several computer languages for a range of different uses and purposes. However, for the experiment reported here *Area* has been completely rewritten. Today it is a simpler version that runs completely on the client side from a standard browser. *Area* is free software.

In presenting this alternative visualization proposal, we have selected the contents of the journal *Information Research* to serve as our corpus of texts on which we demonstrate the tool's visualization and exploration capacities. To do so, we replicated these contents on a standalone server, where *Area* is presented as an alternative interface to that of the *Information Research* journal, yet emulating all its capabilities and adding additional ones (see Table 1) (Nualart, 2014a).

We have chosen the contents of *Information Research* for two reasons. On the one hand, it serves as a good example of an open access journal with the collection being published online under a creative commons license and, on the other, academic papers represent a controlled collection of texts with a similar language register, structure and length, which gives the collection a homogeneous shape. All the codes related to this experiment, as well as the *Area* software itself, can be downloaded from the GitHub repository (Nualart, 2014b).

3.1. Area's visualization capacity

Generally speaking, *Area* is an architectural proposal in which we articulate, in a single structure, the two main systems facilitating the location of information in digital contexts, i.e., the navigation system and the search system. These two systems are present in most contexts, which is a guarantee that users are fully familiar with them and that additional specific instructions are not needed for them to use *Area* efficiently and comfortably.

The combination of navigation and searching interfaces has been presented as an advantage. Morville (2006) argue that the search and navigation systems may be articulated through meta-data indexing; thus, in the web interface, the navigation system categories are, in turn, the fields of the search system. In her classic book, Hearst (2009) studies a list of methods where interfaces support browsing as part of the search process, and shows how each method depends on the size of the data represented. Finally, the recent work called LogSpider (Stange et al., 2014) proposes a method to browse big log files that also combines browsing and searching in a single graphical overview.

Area represents two of the eligible properties simultaneously. The first property is represented graphically as blocks. These form a grid of blocks that contain the items in the collection, depicted as small squares. The second property is the color representation of each item (see Fig. 1). This particular architectural structure provides the tool with a series of capabilities for locating and

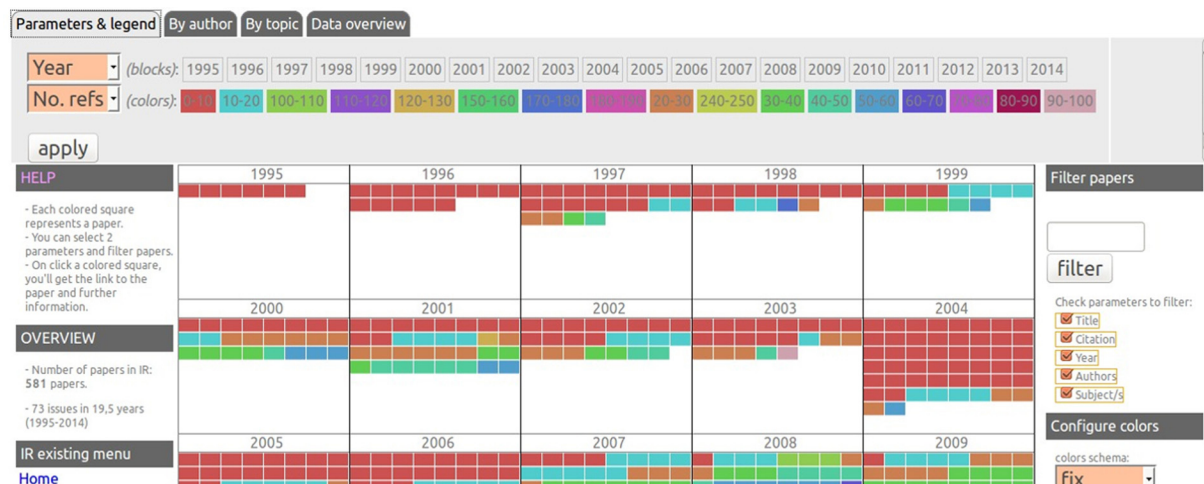


Fig. 1. Screenshot of *Area* interface. The first eligible property is “year” of publication, represented by twenty-five blocks. The second eligible property is “no. of references/papers”, grouped in seventeen categories and is represented by a different color.

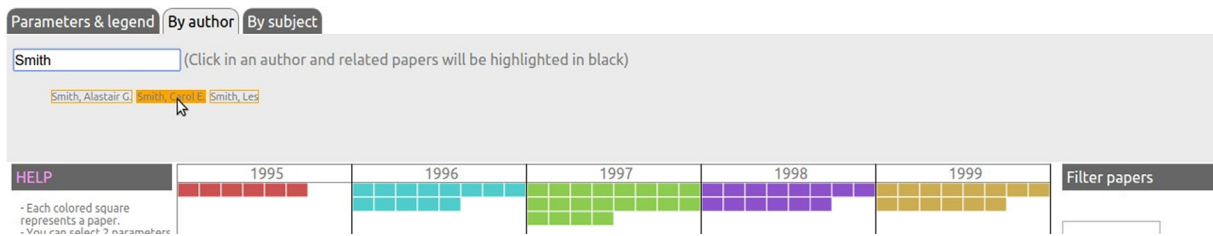


Fig. 2. Detail of filter “by author”.

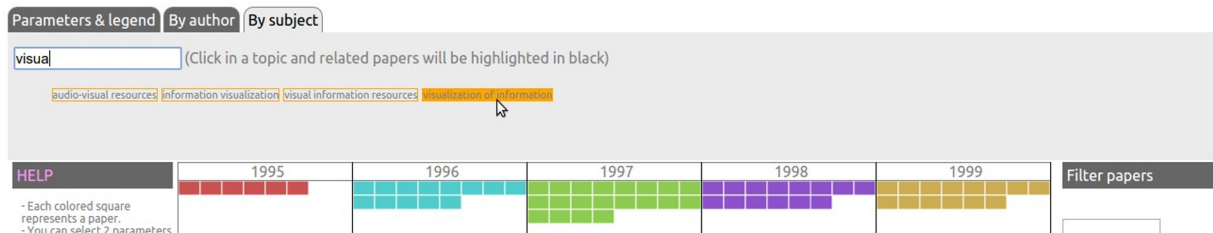


Fig. 3. Detail of filter “by subject”.

visualizing the information contained in the collection that makes up the web page of the journal or digital library.

First, the system can browse the collection and simultaneously select two of the attributes of each document in the collection: the year of publication, the volume in which it appears, the issue in which it was published, the number of references per article and the language in which the article is written. The application of this double selection process generates a two-dimensional representation in which all (not just part) of the collection of documents managed on the web page of the journal or in the digital library is depicted, unlike classical navigation and search systems. This presentation allows us to visualize information about the collection, such as the volume of the collection referred to, the way in which the volume or issues are distributed throughout the year, the annual variation in the number of references included in the documents and the distribution of articles by languages. These indicators are not available in classical systems.

By clicking on one of the rectangles (representing a document) in the grid, a central window opens showing all the available bibliographic information (title, author, volume, number of references, etc.) about that selected paper. The system allows 25 combinations of “eligible properties” (5×5), of which twenty relate two different properties (bivariates) and five represent the collection in terms of a single attribute (univariates), where blocks and colors coincide. Fig. 1 shows the entire collection of documents from *Information Research* using as our criteria the year and the number of references. Each block corresponds to a year and each rectangle corresponds to a document colored according to the number of references that it includes.

Second, once the collection has been presented in terms of the combination of criteria or parameters, the system allows us to apply a series of filters to locate documents that can help the user meet her information needs. The documents corresponding to the filters are highlighted in black. Specifically, three different types of filter are available: author, subject and manual with field selection.

If we click the tab marked “by author” tab (top left), we can write the name of the author of the documents we seek or choose the author from the list of all authors that have published in the journal. This second option should be understood as a system query-builder (Fig. 2).

If we click the tab marked “by subject” (top left), we can write the subject of the documents we seek or choose the subject from the list of all subjects dealt with by the documents in the collection.

This second option should, once more, be understood as a system query-builder (Fig. 3).

The manual filter – “filter paper” (in the right-hand column and mid-zone) – allows a text to be filtered by the attributes or parameters of the document, namely, title, citation, year, authors or subject’s (Fig. 4). The user can choose which of these fields they want to filter for. If more than one filter is selected, the OR operator functions between them. If the author filter is selected, we can also filter by the university to which the author is affiliated or the city in which the author lives.

It should be pointed out that the first two filters (“by author” and “by subject”) are not cumulative, so that every time we write something in the corresponding box this overrides the previous filter. Once one of these two filters is completed, if we change the attributes, the filter is maintained. Moreover, once we have used one of those two filters, we can see what we have typed using the query builder (clicking) as it will appear written in the manual filter box.

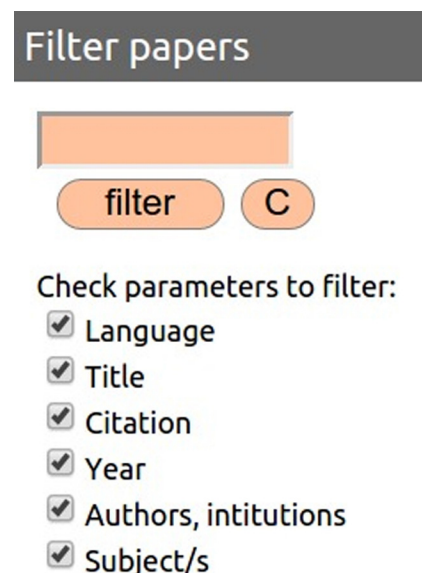


Fig. 4. Detail of manual filter.

Third, *Area* allows us to customize the visualization by giving users the possibility of varying the colors (fixed, random or gradient mode) and thus overcome any potential problems of color-blindness that users might suffer from. It should also be stressed that it incorporates (left-hand column and mid-zone) a support text which explains how to use the tool and an overview of the data in the collection making up the journal or digital library web page.

Fourth, *Area* also includes the original location systems available on the *Information Research* website. Thus, the filters offered by the tool can be understood as a complement to the Google and Atomz searches offered by the *Information Research* website.

Finally, by incorporating a grid that grows in function of the size of the collection, and not depending on other systems such as 3-D or clusters, it avoids the potential visual overlapping of information and the production of visual noise when representing large quantities of documents. *Area*, as specified in the technical description, is recommended for collections of up to 50,000 items (Nualart, 2014c).

3.2. Technical description

Area is a simple, small application coded in javascript, which uses the libraries jquery and D3, HTML, and CSS. The data files are stored in JSON format and the application is accessible with a modern browser. When visiting the *Area* website the client can download all the necessary files to run the application entirely on the client side.

The implementation of this application faces two main constraints: the number of items in the dataset and the dataset size. The first of these is related to screen resolution while the second is related to the size of RAM memory available on the client side. Performance tests conducted³¹ suggest the use of collections that do not exceed fifty thousand items.

Area represents the metadata of a collection of items, allowing filtering and the exploration of the contents of each item. Each time the application and the data files are downloaded, the properties from the metadata schema are analyzed. In those cases in which the number of possible values of a property is not greater than a configured value, then the group of eligible properties is added.

Area was tested in 2014 on desktops, laptops, mobile phones and tablets. All were found to offer good interface responsiveness. However, small screens need to use scroll and zoom in order to provide the same experience as that on larger screens.

4. User evaluation test

To gain a better understanding of the potential of the visual exploration and search of text collections with the *Area* tool, we undertook semi-structured interviews, and a web-based survey.

During the process of development of the interface, the early versions have been assessed by experts through semi-structured interviews. We conducted five interviews of experts in the field of graphic design, media art, and hardware/software development. Five participants is a good number for catching design errors in the early stages of software development. It is a technique that provides insight, not validation (Bevan et al., 2003).

The semi-structured interviews have helped us to simplify the interface design, and to focus on overcoming every initial existing journal web page feature in order to show that all classic features can be undertaken in a visual way.

The aim of the survey was to compare the text-based website with *Area* for the presentation of collections of texts, specifically, scientific papers. To this end, we addressed the following questions: *Area* users able to detect the new features? Do users still prefer or require access to the existing presentation? Are users able to understand the new features? Do users feel confident and positive about using the new features?

The design of the experiment is based on the established technology acceptance model (TAM) (Davis et al., 1989), and the task technology fit (TTF) (Goodhue, 1995). TAM seeks to understand why people accept or reject information technologies, whereas TTF says that technologies will be used if, and only if, their available functionalities support the user's activities. As such, the focus is on the match between the user's task needs and the available functionalities of a given technology. The questions have been designed following Taylor-Powell and Marshall (1996).

In the rest of this section we explain the data collection process: choice, download and storage. Then we describe the demographics of the participants. Finally, we explain in detail the content of the questionnaire administered to the users. In the section the follows we discuss the results of the evaluation.

4.1. Data collection

We used the collection of papers in *Information Research* (IR), edited by Prof. T.D. Wilson (<http://www.informationr.net/ir>). It has been published since 1995, and as of November 2014 the journal has published 592 papers, in 74 quarterly issues, and 19 yearly volumes.

We selected the contents of *Information Research* to provide the corpus of texts for this experiment for the two main reasons discussed above, namely its status as an open access journal, published online under a creative commons license and, because its academic papers constitute a controlled collection of texts with a similar language register, structure and length, giving the collection a homogeneous shape.

In designing *Area* we sought to provide most of the features that the existing *Information Research* website offers. Indeed, for some features *Area* redirects the user to the existing services on the

Table 2
Metadata properties: list and details.

Metadata properties	Type	In which page of the journal is this data?	Function (no. of different values)
Volume	Integer	Paper page + issue page	Eligible property (nineteen volumes)
Issue	Integer	Paper page + issue page	Eligible property (seventy-four issues and four values)
Year	Integer	Paper page + issue page	Eligible property (twenty years)
Number of references (grouped)	Integer	Paper page	Eligible property (seventeen groups)
Language	String	Paper page + issue page	Eligible property & searchable (three languages: English, Spanish, Portuguese)
Title	String	Paper page + issue page	Searchable (592 values)
Authors, institution/country	String	Paper page	Searchable (582 values)
Citation	String	Paper page	Searchable (592 values)
Paper URL	URL string	Paper page + issue page	External link (592 values)
Issue URL	URL string	Paper page + issue page	External link (74 values)
Number of references	Integer	Paper page	Property (101 values)
Paper subjects	String	By-subject page	Property (400 subjects)
Individual author names	String	By-author page	Property (895 authors)

Table 3

Task questionnaire.

How many papers have been published in the journal since the first issue?			How is the term "visualization" distributed in the history of the journal?		
Answer	Count	Percentage	Answer	Count	Percentage
IR existing interface	2	5.41%	IR existing interface	1	2.70%
IR Area interface	31	83.78%	IR Area interface	31	83.78%
No difference	4	10.81%	No difference	5	13.51%
No answer	0	0.00%	No answer	0	0.00%
How many papers talk about visualization?			When exploring papers of the journal website: do you have a better overview of the journal using the existing interface or the Area interface?		
Answer	Count	Percentage	Answer	Count	Percentage
IR existing interface	3	8.11%	IR existing interface	6	16.22%
IR Area interface	24	64.86%	IR Area interface	30	81.08%
No difference	10	27.03%	No difference	1	2.70%
No answer	0	0.00%	No answer	0	0.00%
Understanding the topics and themes of the journal			Finding papers related to your personal interests		
Answer	Count	Percentage	Answer	Count	Percentage
IR existing interface	4	10.81%	IR existing interface	3	8.11%
IR Area interface	31	83.78%	IR Area interface	30	81.08%
No difference	2	5.41%	No difference	4	10.81%
No answer	0	0.00%	No answer	0	0.00%
Exploring new topics and discovering new research in this field					
Answer	Count	Percentage			
IR existing interface	5	13.51%			
IR Area interface	29	78.38%			
No difference	3	8.11%			
No answer	0	0.00%			

website. This is the case of Atomz search and the domain-restricted Google search. Other features have been improved in *Area*, specifically, exploring the collection by year, by language, by number of references per paper, by issue and by volume, and exploring by subject and by author.

To obtain the data collection we harvested the contents from the journal's website. Papers have been published in different versions of HTML, reflecting the evolution of the markup language since 1995 and changes dictated by the publishers in the structure of the pages. We customized the spiders to the non-homogeneous HTML structure of the corpus.

After cleaning the data and adding HTML entities for all special characters, above all for authors' names, we selected several metadata properties for each paper. See Table 2 for a detailed list.

In line with the conditions described above, five properties were labeled as being eligible: volume, issue number, year, grouped number of references/papers, and language. The remaining properties (eight) were: title, authors with institutions and countries, citation, paper URL, issue URL, number of references, paper subjects, and individual author names.

This valuable metadata from the contents of *Information Research* were stored in JSON files, and like the rest of the code, have been published under free licenses to allow others to reuse them.

4.2. Questionnaire description

Online questionnaires are the most frequently employed method for collecting quantitative data from users for statistical analysis. Questionnaires allow the participation of an unlimited number of people and can be used to gather data about users' knowledge, beliefs, attitudes, and behaviors³⁴. Online questionnaires also make it easier to protect the privacy of participants.

The questionnaire comprised fourteen questions. Seven demographic questions and seven specific questions compare the tasks and features of the *Information Research* website and the *Area* website.

Eligible respondents of the questionnaire were any potential visitors of an academic journal. Initially, participants were invited to visit the existing *Information Research* website and the *Area* website in order to familiarize themselves with them and so as to be able to answer the questions. In order to find participants, open calls were sent out using mailing lists of PhD and Master's students.

5. Results

The questionnaire was answered by forty-four respondents, with thirty-seven completing all the questions. Therefore this evaluation study shows the user acceptance of a new tool (*Area*) to and search journal contents. Here we are not conducting a usability evaluation, otherwise it would be necessary a much bigger participation and a more specific set of questions.

One out of three respondents were women and seven out of ten were between thirty and fifty years of age. All the participants said they had either a good, very good or expert technical knowledge of computers in approximately equal proportions. In line with this, seven out of ten of the participants use web browsers several times a day.

The attitude of the participants to the new features found on the websites was positive: they like to find new features sometimes (56.82%) or often (15.91%). Other answers were: no opinion (18.18%), and rarely (4.55%). In contrast, almost half of the participants said they were happy (45.45%) with the information tools and interfaces they use. Finally, more than half of the participants (56.41%) have published scientific papers, and three out of four read scientific papers on quite a regular basis.

We asked participants to compare several tasks completed with the journal's existing interface, on the one hand, and with *Area*'s interface, on the other. To answer the questions we encouraged participants to visit both sites and to familiarize themselves with their interfaces before they started to complete the questionnaire. For all seven tasks, users preferred the new interface. In six out of the seven, participants preferred *Area* for solving the proposed tasks in 80% of cases.

6. Discussion

Taking into account the need for a larger evaluation study in the future, the results of the questionnaire do indicate to us that *Area* is seen to be a promising tool that gives users a set of features that go beyond the conventional tools available on the website of the *Information Research* journal.

Area was preferred by 80% of the users for completing the following tasks:

- Verifying the number of papers making up the collection.
- Identifying the number of papers addressing a specific subject and their distribution in time.
- Obtaining an overview of the collection.
- Understanding the subjects addressed by the journal.
- Finding papers related to a user's interests.

And 64.86% of users preferred it for:

- Exploring new topics and discovering new research in a specific field.

In the case of the following functions: explore by issue as a list of papers, search with Atomz, and search with Google, *Area* redirects users to the resources on the journal's web page.

However, *Area* betters the classical visualization tool (included in the interface of most journals and digital libraries) in several of its features (Table 3). On the one hand, it incorporates new visualization features that are not available in the classical proposal. For example, it allows the user to visualize the whole collection in different ways depending on the two properties and filters selected, and not just as a subset of the whole as in classical systems. We have named this new function: *multiple overviews of the collection*. Although, *Area* provides rapid access to the quantitative characteristics of the collection (numbers of papers, issues, volumes, years, etc.), a function named *numerical overview*, it is preferred for less users that the other tasks in the question: "How many papers talk about visualization?". The *Information Research* journal home page offers a domain specific Google search box. If you use this Google search box it will not give you the number of papers, as it is in the question, but probably, participants feel confident in using Google. *Area* also shows the user how a subject is distributed during the history of the journal as it allows filtered papers to be marked during exploration. A feature we have named *topic distribution*. And, finally, *Area* allows the user to explore papers by language and to see the evolution in this language, since language is an eligible property and it can be represented in combination with the other properties. We have named this new function: *Explore by language*.

Area also improves certain functions that already exist in the classical version. For example, it improves the explore by year, issue and volume function by allowing multiple representations and evolution visualization. It also improves the functions of explore "by author" and explore "by subject" by allowing filter-by-typing. Finally, *Area* improves the function of identifying how many papers talk about a subject? By showing the papers and their context.

7. Conclusions

These new visualization functions and the outcomes recorded allow us to draw a number of conclusions.

The evaluation study shows that the user acceptance of a new tool, *Area*, is positive: users detect and understand new features; users prefer or require new ways of presenting information; and users feel confident and positive about using the new features. However, our evaluation study is clearly not a usability evaluation,

in which it would be necessary to have a much larger sample of survey participants and a more specific set of questions.

The simplicity and economy of the *Area* prototype should pave the way for the widespread introduction of these visualization tools in the portals and websites of journals and digital libraries. The fact that *Area* is not implemented as a page which is independent of the basic search interface means that it is not perceived by users as a secondary tool; nor does the prototype present a high level of abstraction and conceptualization that means its use is not very intuitive for users. Similarly, *Area*, by basing its visualization power on the metadata file, is a non-intrusive system that only needs to be accessible from any point in the network and, once downloaded locally, it allows interaction without an Internet connection. Unlike other prototypes that have been implemented only with small collections of documents and in highly controlled experimental conditions, *Area* has been implemented in a real world context with the entire collection of documents from a journal (not just with a subset of retrieved documents). Therefore, the user-satisfaction results reported here cannot be dismissed on the grounds of their having been obtained with a limited collection or a limited number of documents. Finally, it should be stressed that *Area* is a free licensed tool that is readily implemented which, unlike other more abstract and expensive prototypes, facilitates its implementation in journal and digital library sites.

Acknowledgements

This work is part of the Project iActive Audiences and Journalism. Interactivity, Web Integration and Findability of Journalistic Information. CSO2012-39518-C04-02. National Plan for R + D + i, Spanish Ministry of Economy and Competitiveness.

References

- Keith, Andrews, Gütl, Christian, Moser, Josef, Sabol, Vedran, & Lackner, Wilfried. (2015). Search result visualisation with xfind. In *In User Interfaces to Data Intensive Systems, 2001. UIDIS 2001 Proceedings. Second International Workshop on*, (pp. 50–58). IEEE, 2001
- Andrews, Keith, Kienreich, Wolfgang, Sabol, Vedran, Becker, Jutta, Droschl, Georg, Kappe, Frank, Granitzer, Michael, Auer, Peter, & Tochtermann, Klaus. (2002). *The infosky visual explorer: exploiting hierarchical structure and document similarities*. *Inf. Visualization*, 1(3–4), 166–181.
- Baeza-Yates, Ricardo, 2011. Tendencias en recuperación de información en la web. *Bid*, no. 27, 1–4.
- Bates, Marcia J. (1989). *The design of browsing and berrypicking techniques for the online search interface*. *Online Rev.*, 13(5), 407–424.
- Bauer, Sabine. (2014). *Interactive Visualizations for Search Processes*. 5th IEEE Germany Student Conference. University of Passau.
- Bevan Nigel, et al. (2003). *The Magic Number 5: Is it Enough for Web Testing? CHI'03 Extended Abstracts on Human Factors in Computing Systems*. ACM.
- Chalmers, Matthew, & Paul, Chitson. (1992). *Bead: explorations in information visualization*. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, 330–337. ACM, 1992.
- Cugini, John V., Laskowski, Sharon, & Sebrechts, Marc. (2000). *Design of 3D visualization of search results: evolution and evaluation*. *Electron. Imaging*, 198–210.
- Davis, Fred D., Bagozzi, Richard P., & Warshaw, Paul R. (1989). *User acceptance of computer technology: a comparison of two theoretical models*. *Manage. Sci.*, 35(8), 982–1003.
- Egan, Dennis Egan, Remde, Joel R., Gomez, Louis M., Landauer, Thomas K., Eberhardt, Jennifer, & Lochbaum, Carol C. (1989). *Formative design evaluation of superbook*. *ACM Trans. Inf. Syst. (TOIS)*, 7(1), 30–57.
- Fox, Edward A., Hix, Deborah, Nowell, Lucy T., Brueni, Dennis J., Wake, William C., Heath, Lenwood S., & Rao, Durgesh. (1993). *Users, user interfaces, and objects: envision, a digital library*. *J. Am. Soc. Inf. Sci.*, 44(8), 480–491.
- Goodhue, Dale L. (1995). *Understanding user evaluations of information systems*. *Manage. Sci.*, 41(12), 1827–1844.
- Hearst, Marti A., & Karadi, Chandu. (1997). *Cat-a-cone: an interactive interface for specifying searches and viewing retrieval results using a large category hierarchy*. *ACM SIGIR Forum*, 31, 246–255.
- Hearst, Marti A. (1995). *Tilebars: visualization of term distribution information in full text information access*. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 59–66.
- Hearst, Marti. (2009). *Search User Interfaces*. Cambridge University Press.

- Hienert, Daniel, Sawitzki, Frank, Schaer, Philipp, & Mayr, Philipp. (2012). *Integrating Interactive Visualizations in the Search Process of Digital Libraries and IR Systems*. In *Advances in Information Retrieval*. Springer.
- Hoeber, Orland, & Yang, Xue Dong. (2006). *A comparative user study of web search interfaces: HotMap, concept highlighter, and Google*. En *web intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on*, 866–874.
- Jones, Steve. (1998). *Graphical query specification and dynamic result previews for a digital library*. En *Proceedings of the 11th annual ACM symposium on User interface software and technology*, 143–151.
- Kim, Beomjin, Scott, Jon, Kim, Seung, Eun, 2011. Exploring digital libraries through visual interfaces.
- Lam, Heidi, & Baudisch, Patrick. (2005). *Summary thumbnails: readable overviews for small screen web browsers*. En *Proceedings of the SIGCHI conference on Human factors in computing systems*, 681–690.
- Morville, Peter. 2006. Information architecture 3.0. Semantics Studios 29.
- Morville, Peter, & Rosenfeld, Louis. (2007). *Information Architecture*. Sebastopol: O'Reilly Media, Inc.
- Nualart, Jaume, 2014a. Area for Information Research, (accessed 01.03.15.) <<http://research.nualart.cat/area-ir/>>.
- Nualart, Jaume, 2014b. Area repository. (accessed 01.03.15.) <<https://github.com/jaumet/Area>>.
- Nualart, Jaume, 2014c. Area stress. (accessed 01.03.15.) <<http://research.nualart.cat/area-stress/>>.
- Nualart, Jaume, & Pérez-Montoro, Mario. (2013). *Texty, a visualization tool to aid selection of texts from search outputs*. *Inf. Res.*, 18(2)
- Nualart, Jaume, Pérez-Montoro, Mario, Whitelaw, Mitchell, 2014. How we draw texts: A review of approaches to text visualization and exploration. *El profesional de la información* 23, (3) 221–235.
- Pérez-Montoro, Mario, 2010. Arquitectura de la información en entornos web. *El profesional de la información* 19, (4) 333–338.
- Robertson, George G., Mackinlay, Jock D., Card, Stuart K., 1991. Cone trees: animated 3D visualizations of hierarchical information. En *Proceedings of the SIGCHI conference on Human factors in computing systems*, 189–194.
- Schatz, Bruce R., Johnson, Eric H., Cochrane, Pauline A., Chen, Hsinchun, 1996. Interactive term suggestion for users of digital libraries: using subject thesauri and co-occurrence lists for information retrieval. En *Proceedings of the first ACM international conference on Digital libraries*, 126–133.
- Shneiderman, Ben, Feldman, David, Rose, Anne, Grau, Xavier Ferré, 2000. Visualizing digital library search results with categorical and hierarchical axes. En *Proceedings of the fifth ACM conference on Digital libraries*, 57–66.
- Stange, J. E., Dörk, M., Landstorfer, J., & Wettach, R. (2014). *Visual filter: graphical exploration of network security log files*. In *Proceedings of the Eleventh Workshop on Visualization for Cyber Security ACM*, (pp. 41–48). November.
- Taylor-Powell, Ellen, & Marshall, Mary Gladys. (1996). *Questionnaire Design Asking Questions With a Purpose*. University of Wisconsin-Extension Cooperative Extension Service.
- van Hoek, Wilko and Mayr, Philipp, 2014. Is Evaluating Visual Search Interfaces in Digital Libraries Still an Issue? arXiv preprint arXiv:1408.5001.
- Weiss-Lijn, Mischa, McDonnell, Janet T., & Leslie, James. (2001). *Supporting document use through interactive visualization of metadata*. En *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*.
- Wong, William, Chen, Raymond, Kodagoda, Neesha, Rooney, Chris, Xu, Kai, 2011. INVISQUE: intuitive information exploration through interactive visualization. En *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, 311–316.
- Woodruff, Allison, Faulring, Andrew, Rosenholtz, Ruth, Morrisson, Julie, & Pirolli, Peter. (2001). *Using thumbnails to search the Web*. En *Proceedings of the SIGCHI conference on Human factors in computing systems*, 198–205.

B. Appendix: Forty-nine text visualization approaches

Single text visualization

Whole text visualization

1 ▸ Novel Views: Les Misérables - Radial Word Connections by Jeff Clark (2013)

- data: Novel "Les Misérables"
- method: Radial text-line
- description: "A word used in multiple places in a text can be interpreted as a connection between those locations. Depending on the word itself the connection could be in terms of character, setting, activity, mood, or other aspects of the text."

[<http://neoformix.com/2013/NovelViews.html>]

NOVEL VIEWS - Les Misérables - Word Connections

Radial Word Connections

A word used in multiple places in a text can be interpreted as a connection between those locations. Depending on the word itself the connection could be in terms of character, setting, activity, mood, or other aspects of the text. This graphic shows, for the novel Les Misérables, a number of these word connections.

The 365 chapters of the text are shown with small segments on the inner ring of the circle with the first chapter appearing at the top and proceeding clockwise from there. The outer ring shows how the chapters are grouped into books of the novel and the book titles are shown as well. The words in the middle are connected using lines of the same color to the chapters where they are used.

This small example below shows that the author devoted a book to the battle of Waterloo at the beginning of the second volume and that there were a few scattered references elsewhere. Similarly, we can see with the blue that there is another book entirely about slang.



Jeff Clark - neoformix.com - © 2013

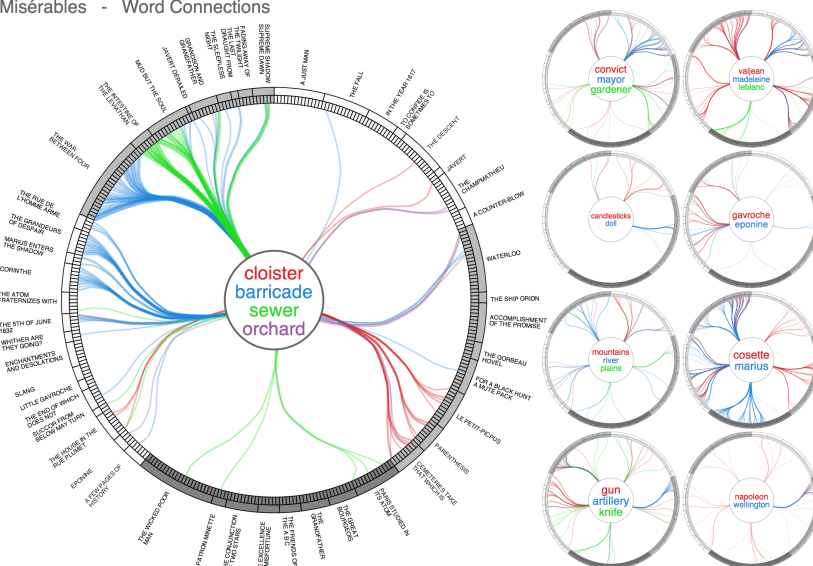


Figure 0.1: Novel Views: Les Misérables - Radial Word Connections by Jeff Clark (2013)

2▷ Novel Views: Les Misérables - Character Mentions by Jeff Clark (2013)

- data: Novel "Les Misérables".
- method: Horizontal multi text-line.
- description: "it shows where the names of the primary characters are mentioned within the text. Click on any of these images to see larger versions."

[<http://neoformix.com/2013/NovelViews.html>]

NOVEL VIEWS - Les Misérables - Character Mentions

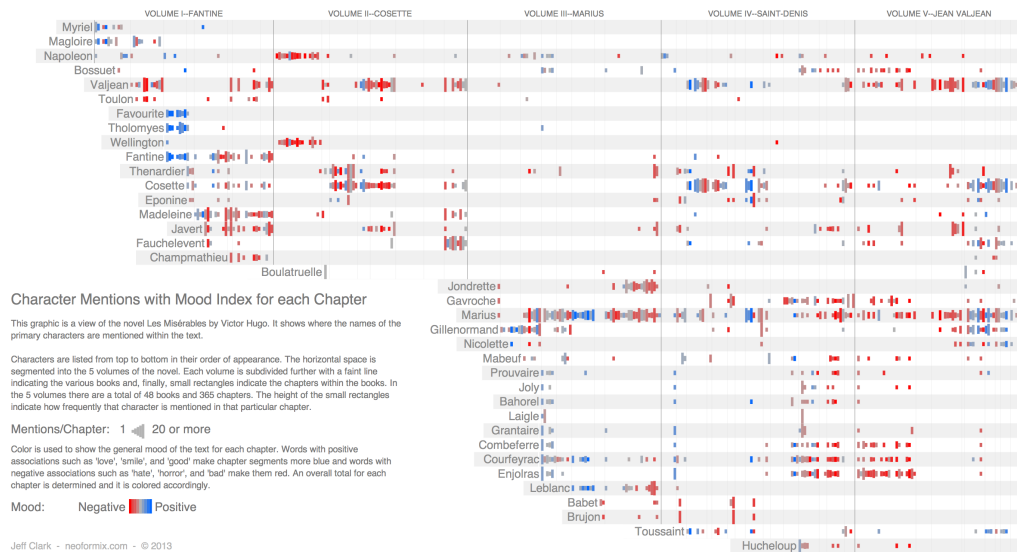


Figure 0.2: Novel Views: Les Misérables - Character Mentions by Jeff Clark (2013)

3 ▷ Poem Viewer - (on process) 2013 project Imagery Lens for Visualizing Text Corpora - Oxford e-Research Centre at the University of Oxford by Katharine Coles, Min Chen, Alfie Abdul-Rahman, Chris Johnson, Julie Lein, Eamonn Maguire, Miriah Meyer, and Martin Wynne. (2013)

- data: Poems.
- method: Advanced Text-line.
- description: Poem Viewer is an experimental service for the exploration and analysis of poetry through visualization. Poem Viewer is part of an on-going research project and is currently a work in progress.

[<http://ovii.oerc.ox.ac.uk/PoemVis/>]

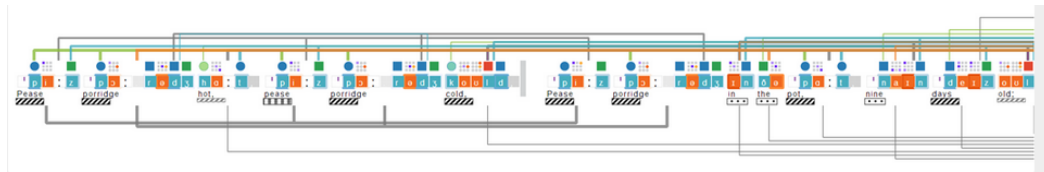


Figure 0.3: Poem Viewer - (on process) 2013 project Imagery Lens for Visualizing Text Corpora - Oxford e-Research Centre at the University of Oxford by Katharine Coles, Min Chen, Alfie Abdul-Rahman, Chris Johnson, Julie Lein, Eamonn Maguire, Miriah Meyer, and Martin Wynne. (2013)

4▷ State of the Union 2011 - Sentence Bar Diagrams by Jeff Clark (2011)

- data: Speech to text. Obama's State of the Union speech 2011 [N/A]
- method: Sentence Bar Diagrams.
- description: "First we have two Sentence Bar Diagrams for the speeches from 2010 and 2011. Sentence Bar diagrams use color coding to show the topic of the various sentences in the text and bar length to show how long the sentences are."

[<http://neoformix.com/2011/SOTU2011.html>]



Figure 0.4: State of the Union 2011 - Sentence Bar Diagrams by Jeff Clark (2011)

5▷ Visualizing Lexical Novelty in Literature by Matthew Hurst (2011)

- data: Literature
- method: Original pixel-block method
- description: Tracking the introduction of new terms in a novel. In the visualization each column represents a chapter and each small block a paragraph of text. The color of the block indicates the % of new words. It is not clear what is the utility of this tool; an specialist in linguistics could say how this visualization can be used.

[http://datamining.typepad.com/data_mining/2011/09/visualizing-lexical-novelty-in-literature.html]

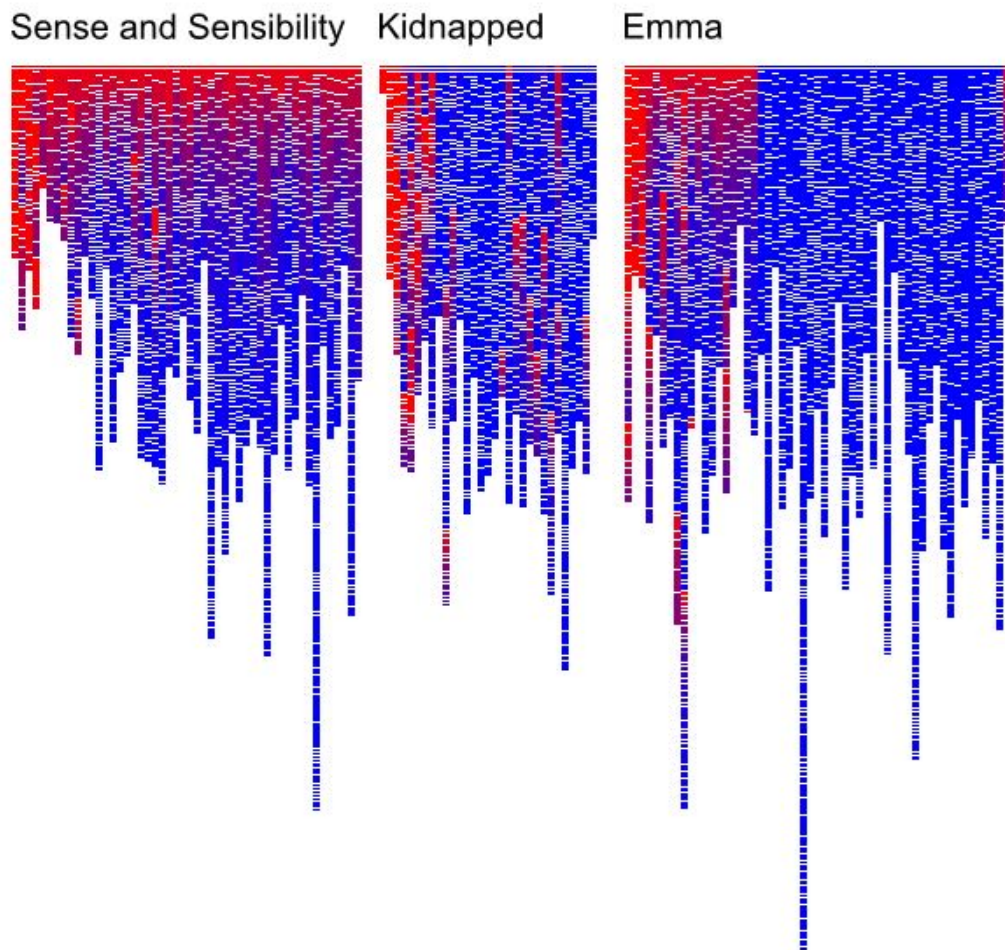


Figure 0.5: Visualizing Lexical Novelty in Literature by Matthew Hurst (2011)

6 ▷ On the Origin of Species: The Preservation of Favoured Traces by Ben Fry (2009)

- data: All the versions of Darwin's On the Origin of Species Book.
- method: Matrix of pixel-blocks representing all the text in chunks. Colors are used to show each edition additions.
- description: We often think of scientific ideas, such as Darwin's theory of evolution, as fixed notions that are accepted as finished. In fact, Darwin's On the Origin of Species evolved over the course of several editions he wrote, edited, and updated during his lifetime. The first English edition was approximately 150,000 words and the sixth is a much larger 190,000 words. In the changes are refinements and shifts in ideas whether increasing the weight of a statement, adding details, or even a change in the idea itself.

[<http://benfry.com/traces/>]

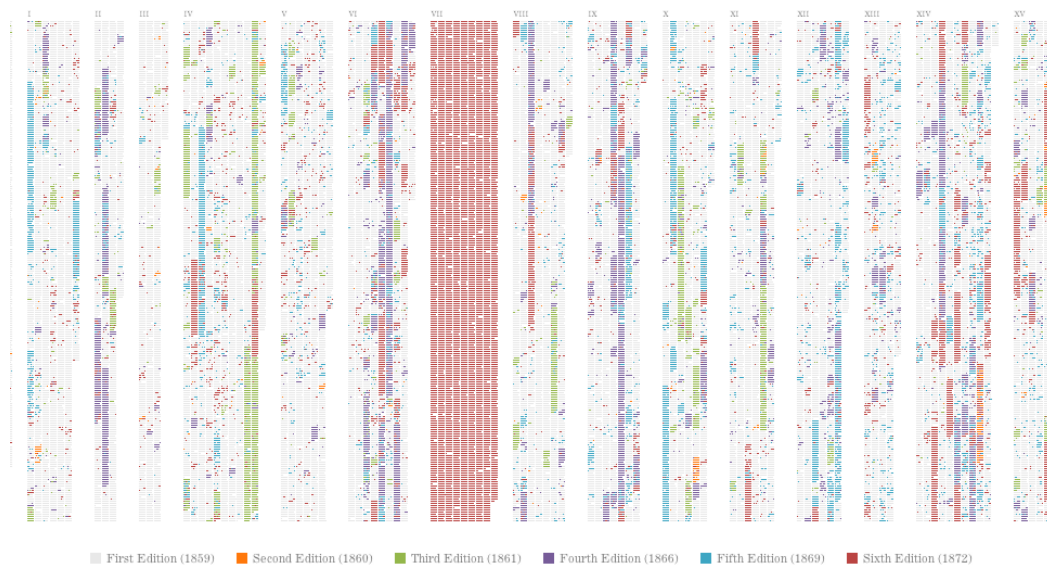


Figure 0.6: On the Origin of Species: The Preservation of Favoured Traces by Ben Fry (2009)

7 ▸ Texty, a visualization tool to aid selection of texts from search outputs by Jaume Nualart (2008)

- data: Ars Electronica Jury Statements (from 1987 to 2007)
- method: Iconic representation of a text
- description: a Texty is an image, an icon that represents the physical distribution of keywords of a text as a flat image. These keywords are grouped in vocabularies, to each of which a color is linked. Texty reveals, the structure, conceptual density and subject matter of a text.
- paper: Nualart, J. Pérez-Montoro, M (2013). "Texty, a visualization tool to aid selection of texts from search outputs" *Information Research*.

[<http://vis.mediaartresearch.at/textass/texty.php>]

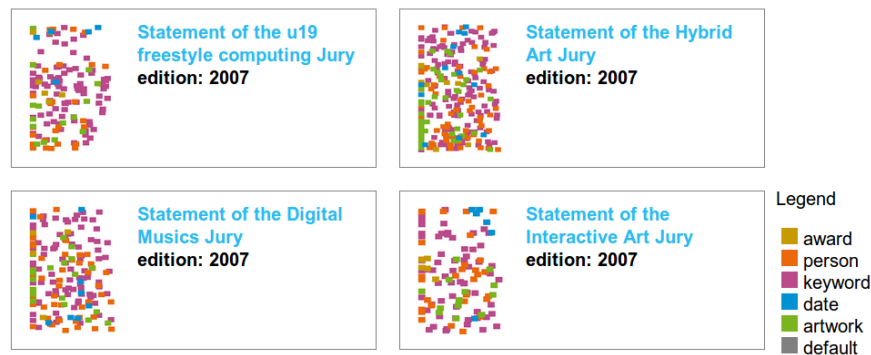


Figure 0.7: Detail of four textys showing texts and the legend from Ars Electronica Jury Statements (2008)

8 ▷ Bible Cross-References by Chris Harrison (2008)

- data: Bible [Large dataset]
- method: Arc graph + bar graph
- description: "The bar graph that runs along the bottom represents all of the chapters in the Bible. Books alternate in color between white and light gray. The length of each bar denotes the number of verses in the chapter. Each of the 63,779 cross references found in the Bible is depicted by a single arc - the color corresponds to the distance between the two chapters, creating a rainbow-like effect."

[<http://www.chrisharrison.net/index.php/Visualizations/BibleViz>]

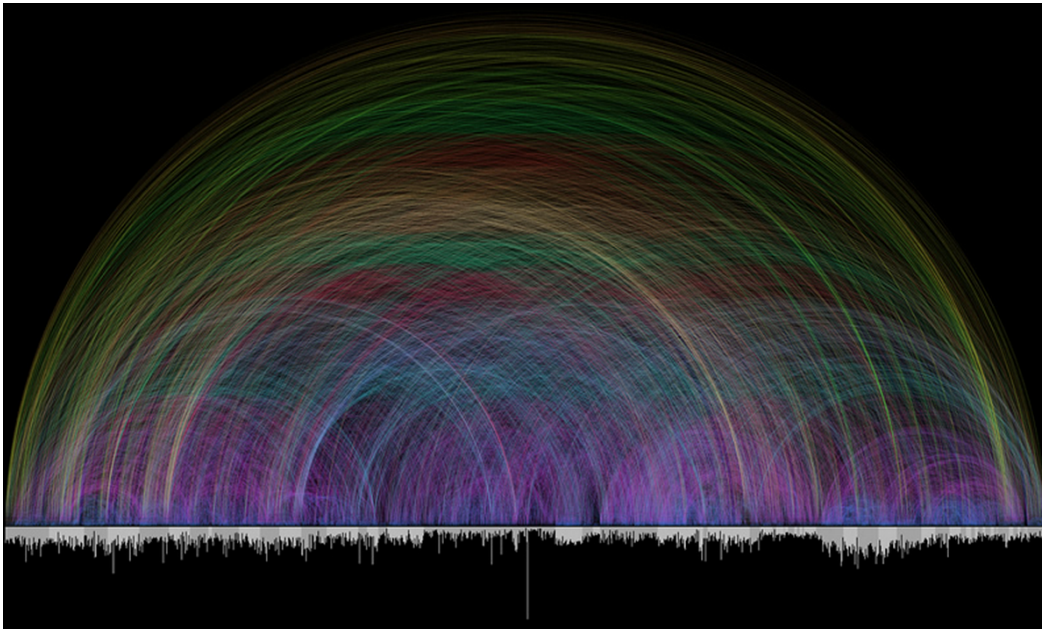


Figure 0.8: Bible Cross-References by Chris Harrison (2008)

9 ▷ Literature fingerprint by Daniel A. Keim and Daniela Oelke (2007)

- data: Any text, from middle to long
- method: Colored squares representing elements of a text in a sequential way.
- description: Colored squares representing elements of a text in a sequential way. Description/Comments: the authors develop visualization techniques to show the results of some linguistic analysis that represent a literature fingerprint, demonstrating text authorship.
- paper: Keim D. A. & Oelke D. (2007). *Literature Fingerprinting: A New Method for Visual Literary Analysis*. In: *Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology (VAST '07)*. IEEE Computer Society, Washington, DC, USA, 115-122. DOI=10.1109/VAST.2007.4389004 <http://dx.doi.org/10.1109/VAST.2007.4389004>

▷ [<http://bib.dbvis.de/uploadedFiles/71.pdf>]

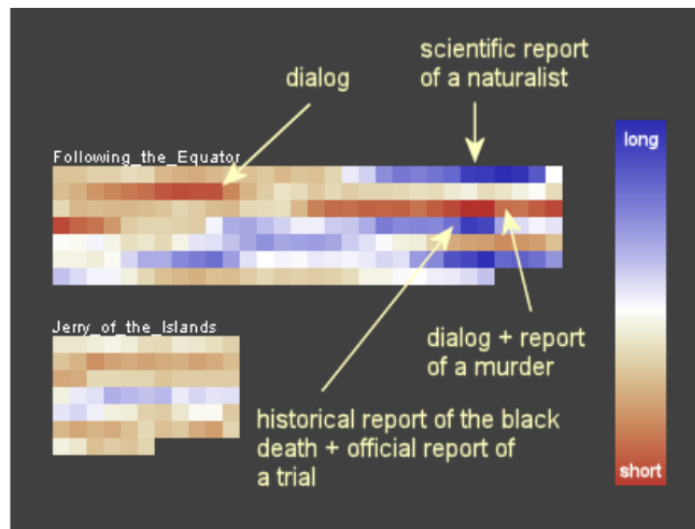


Figure 3: The figure shows the fingerprints of two novels that almost have the same average sentence length. In the detailed view, the different structure of the two novels is revealed. The inhomogeneity of the travelogue *Following the Equator* can be explained with the alternation of dialogs, narrative parts and quoted documents.

Figure 0.9: Sentence length visualization

10▷ History Flow by Fernanda Viégas and Martin Wattenberg (2003)

- data: Wikipedia [Small dataset]
- method: Flow chart time-line
- description: “the history flow application charts the evolution of a document as it is edited by many people using a very simple visualization technique. All text segments contributed by the same editor are marked with a unique color. The diagram shows how some editors produce long lasting content, while others don’t. With enough editors contributing to an article, almost every paragraph or even sentence gets modified in the long run.”
- paper: Viégas, F. B., Wattenberg, M., & Dave, K. (2004, April). *Studying cooperation and conflict between authors with history flow visualizations*. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 575-582). ACM. [http://alumni.media.mit.edu/~fviegas/papers/history_flow.pdf]

[<http://hint.fm/projects/historyflow/>]

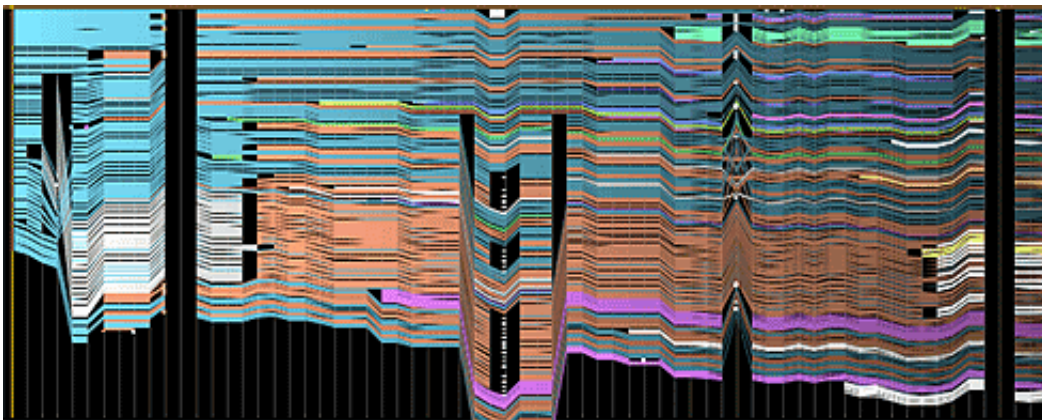


Figure 0.10: History Flow by Fernanda Viégas and Martin Wattenberg researchers at IBM’s Visual Communication Lab (2003)

11▷ Colour-coded chronological sequencing by Joel Deshayes and Peter Stoicheff (2003)

- data: The Sound and the Fury by William Faulkner
- method: Timeline
- description: Case collected by Peter Stoicheff: "This visualization (contributed by Joel Deshayes and Peter Stoicheff) shows a compressed version of the colour-coded "April Seventh, 1928" narrative on the left. The middle bar extracts the narrative of Benjy's present day. The right bar extracts the flashbacks to Caddy's wedding. Although the section seems randomly ordered, within it the present and each flashback reside independently as coherent, chronological sequences."
- paper: Stoicheff, R.P. "Faulkner's Foreign Levy: Macbeth, The Sound and the Fury, and Writerhood." *The Sound and the Fury: a Hypertext Edition*. Ed. Stoicheff, Muri, Deshayes, et al. Updated Mar. 2003. U of Saskatchewan. Accessed 18 Mar. 2003 <<http://www.usask.ca/english/faulkner>>

[http://drc.usask.ca/projects/faulkner/main/benjy_spectrum.htm]

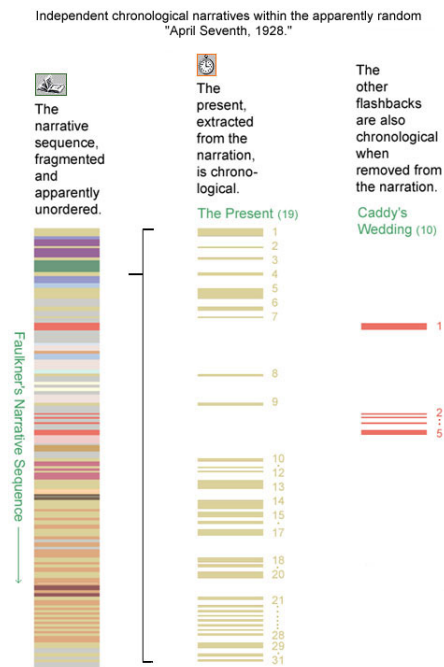


Figure 0.11: Colour-coded chronological sequencing of William Faulkner's novel The Sound and the Fury, by Joel Deshayes and Peter Stoicheff (2003)

12▷ 2-D display of time in the novel by Joel Deshayé (2003)

- data: The Sound and the Fury by William Faulkner
- method: Customized 2-D diagram
- description: Case collected by Peter Stoicheff: "This graph (contributed by Joel Deshayé) shows two dimensions of time in The Sound and the Fury: chronological time, and Faulkner's re-ordering of chronological time into the text's narrative sequence. Faulkner scrambles the chronology not only by flashbacks, but also by the non-linear sequence of the novel's sections. Through this graph it becomes clear that the novel follows a conventional in medias res structure, whose existence is otherwise obscured by more local narrative complexities."
- paper: Stoicheff, R.P. "Faulkner's Foreign Levy: Macbeth, The Sound and the Fury, and Writerhood." *The Sound and the Fury: a Hypertext Edition*. Ed. Stoicheff, Muri, Deshayé, et al. Updated Mar. 2003. U of Saskatchewan. Accessed 18 Mar. 2003 <<http://www.usask.ca/english/faulkner>>

[http://drc.usask.ca/projects/faulkner/main/sf_2d_timegraph.htm]

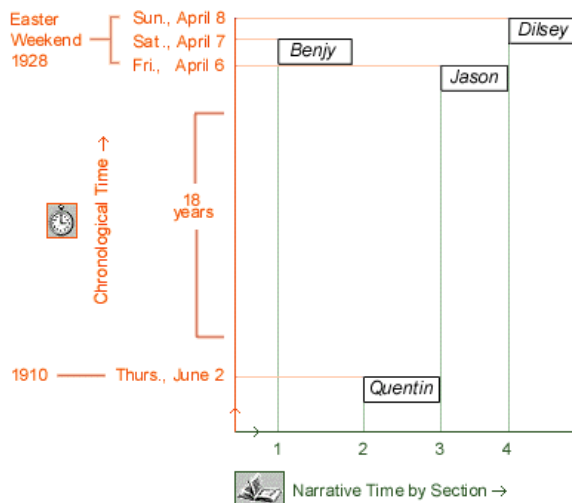


Figure 0.12: 2-D display of time of William Faulkner's novel The Sound and the Fury, by Joel Deshayé and Peter Stoicheff (2003)

13▷ 3-D display of time in the novel by Joel Deshayé (2003)

- data: *The Sound and the Fury* by William Faulkner
- method: Highly customized 3D diagram
- description: Case collected by Peter Stoicheff: "This graph (contributed by Joel Deshayé) represents time in three different dimensions. First, there is the sequence of four sections in the novel - the narrative time or *récit* (shown in green). Second, there is the chronology, *l'histoire* (shown in orange), beginning with the earliest recollected date in the novel - Damuddy's death in 1898. These two dimensions show how the conventional view of linear time can be disrupted by a fictional narrative. Third, there is a representation of the proportion of memories of and flashbacks to the past in each section (shown in blue). This last dimension shows how the novel progresses from an emphasis on the past toward an emphasis on the present."
- paper: Stoicheff, R.P. "Faulkner's Foreign Levy: *Macbeth*, *The Sound and the Fury*, and *Writerhood*." *The Sound and the Fury: a Hypertext Edition*. Ed. Stoicheff, Muri, Deshayé, et al. Updated Mar. 2003. U of Saskatchewan. Accessed 18 Mar. 2003 <<http://www.usask.ca/english/faulkner>>

[http://drc.usask.ca/projects/faulkner/main/sf_3d_timegraph.htm]

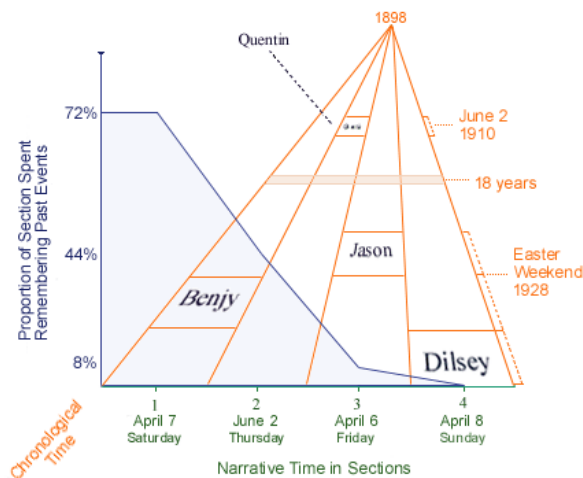


Figure 0.13: 3-D display of time of William Faulkner's novel *The Sound and the Fury*, by Joel Deshayé and Peter Stoicheff (2003)

14 ▷ Arc diagram Wattenberg by Martin Wattenberg (2002)

- data: Any text or string [Small dataset]
- method: Arc diagram
- description: Arc diagram is capable of representing complex patterns of repetition in string data. Arc diagrams improve over previous methods such as dotplots because they scale efficiently for strings that contain many instances of the same subsequence.
- paper: Wattenberg, M. (2002). *Arc diagrams: Visualizing structure in strings*. In *Information Visualization, 2002. INFOVIS 2002. IEEE Symposium on* (pp. 110-116). IEEE.

▷ [[http://domino.watson.ibm.com/cambridge/research.nsf/58bac2a2a6b05a1285256b30005b3953/e2a83c4986332d4785256ca7006cb621/\\$FILE/TR2002-11.pdf](http://domino.watson.ibm.com/cambridge/research.nsf/58bac2a2a6b05a1285256b30005b3953/e2a83c4986332d4785256ca7006cb621/$FILE/TR2002-11.pdf)]



Figure 0.14: Arc diagram by Martin Wattenberg (2002)

15▷ TileBars: Visualization of Term Distribution Information in Full Text Information Access by Marti A. Hearst (1995)

- data: search results from information retrieval systems. Applied to PubMed results.
- method: iconic representation of a text
- description: “TileBars, which provides a compact and informative iconic representation of the documents’ contents with respect to the query terms. The goal is to simultaneously indicate: the relative length of the document, the frequency of the term sets in the document, and the distribution of the term sets with respect to the document and to each other. Each large rectangle indicates a document, and each square within the document represents a TextTile. The darker the tile, the more frequent the term (white indicates 0, black indicates 8 or more instances)”
- paper: Hearst, M. 1995. “TileBars: visualization of term distribution information in full text information access.” Proceedings of the SIGCHI conference on Human <http://dl.acm.org/citation.cfm?id=223912> (March 26, 2013).

[<http://dl.acm.org/citation.cfm?id=223912>]

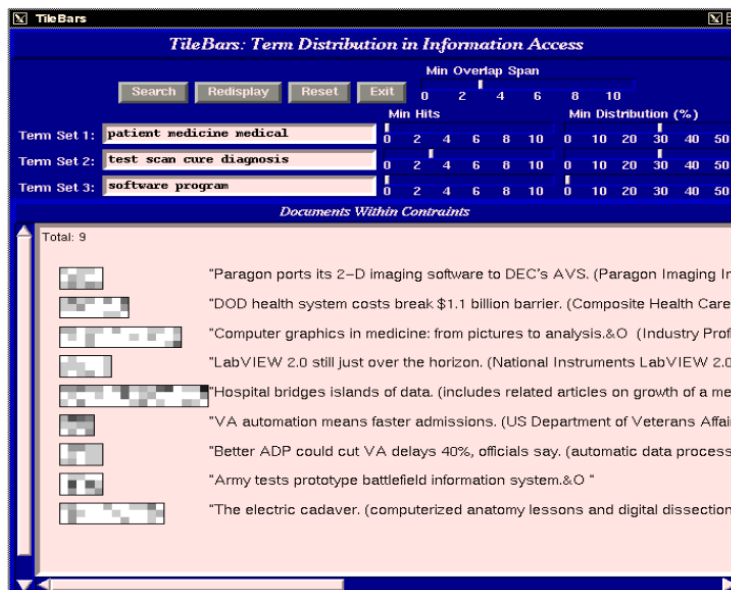


Figure 0.15: TileBar search on (patient medicine medical AND test scan cure diagnosis AND software program) with stricter distribution constraints.

Part text visualization

16 ▸ Novel Views: Les Misérables - Characteristic Verbs by Jeff Clark (2013)

- data: Novel "Les Misérables" [N/A]
- method: Tables + condensed bar graphs
- description: "the verbs used together with character names in a novel can provide a glimpse into the personalities and actions of that character. For the primary people in the novel Les Misérables this graphic illustrates their characteristic verbs."

[<http://neoformix.com/2013/NovelViews.html>]

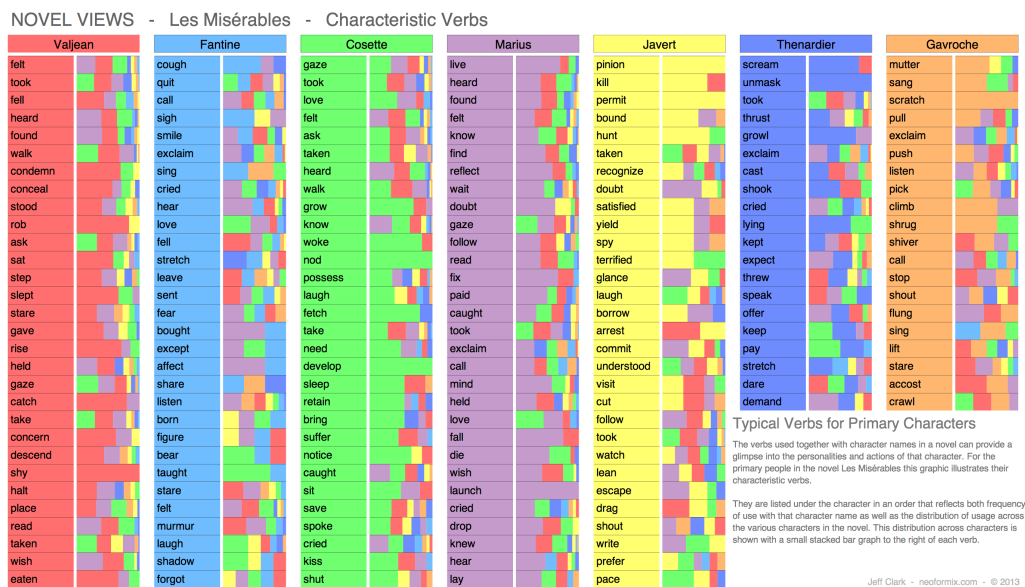


Figure 0.16: Novel Views: Les Misérables - Characteristic Verbs by Jeff Clark (2013)

17 ▷ Wordle by Jonathan Feinberg (2009)

- data: any text
- method: Word cloud
- description: Wordle is a toy for generating “word clouds” from text that you provide. The clouds give greater prominence to words that appear more frequently in the source text.
- paper: *Viegas, F. B., Wattenberg, M., & Feinberg, J. (2009). Participatory visualization with wordle. Visualization and Computer Graphics, IEEE Transactions on, 15(6), 1137-1144.*

▷ [<http://cyber-kap.blogspot.com.au/2011/04/top-10-sites-for-creating-word-clouds.html>]

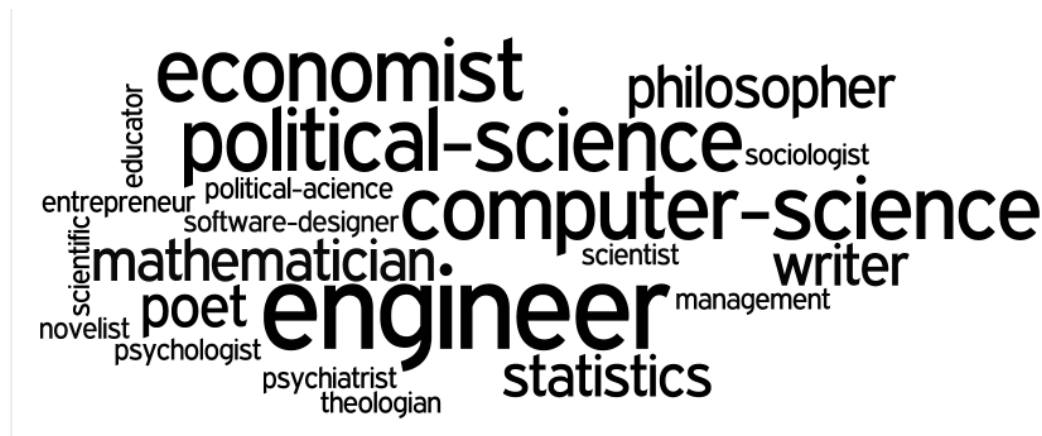


Figure 0.17: Wordle showing the professions of inventors of data visualization tools

18▷ Docuburst by C. Collins, S. Carpendale , and G. Penn (2009)

- data: any text, specially used with books. It is a visualization of Wordnet lexical database.
- method: Radial, space-filling layout of hyponymy (IS-A relation)
- description: DocuBurst is the first visualization of document content which takes advantage of the human-created structure in lexical databases. The authors used an accepted design paradigm to generate visualizations which improve the usability and utility of WordNet as the backbone for document content visualization. A radial, space-filling layout of hyponymy (IS-A relation) is presented with interactive techniques of zoom, filter, and details-on-demand for the task of document visualization. The techniques can be generalized to multiple documents.
- paper: C. Collins, S. Carpendale, and G. Penn, "DocuBurst: Visualizing Document Content Using Language Structure," *Computer Graphics Forum (Proc. of the Eurographics/IEEE-VGTC Symposium on Visualization (EuroVis))*, vol. 28, iss. 3, pp. 1039-1046, 2009.

▷ [<http://vialab.science.uoit.ca/portfolio/docuburst-visualizing-document-content-using-language-structure>]

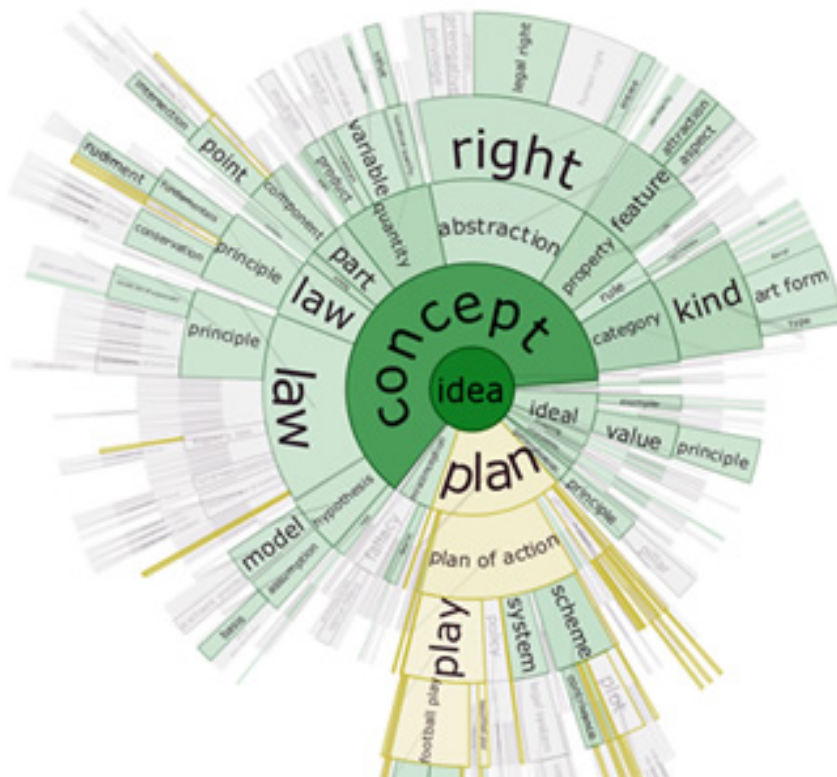


Figure 0.18: Screenshot of docuburst interactive interface.

19▷ Phrase Nets by Frank van Ham, Martin Wattenberg and Fernanda B. Viégas (2009)

- data: Any text. Jane Austen’s novel ”Pride and Prejudice.” [N/A]
- method: Network of sized-words, some connected
- description: “Phrase Nets use a simple form of pattern matching to provide multiple views of the concepts contained a book, speech, or poem. The image below is a word graph made from Jane Austen’s novel ”Pride and Prejudice.” The program has drawn a network of words, where two words are connected if they appear together in a phrase of the form ”X and Y”.”
- paper: *Van Ham, F., Wattenberg, M., & Viégas, F. B. (2009). Mapping text with phrase nets. Visualization and Computer Graphics, IEEE Transactions on, 15(6), 1169-1176.*

[<http://hint.fm/projects/phrasenet/>]

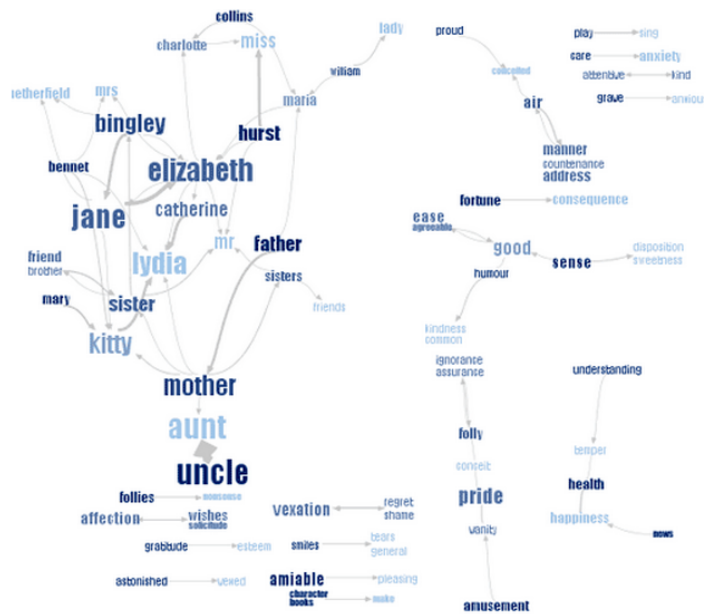


Figure 0.19: Phrase Nets of Jane Austen’s novel ”Pride and Prejudice.” by Frank van Ham, Martin Wattenberg and Fernanda B. Viégas (2009)

20▷ Word Spectrum: Visualizing Google's Bi-Gram Data by Chris Harrison (2008)

- data: The huge Google bi-gram dataset [Large dataset]
- method: Word spectrum
- description: “Using Google’s enormous bigram dataset, I produced a series of visualizations that explore word associations. Each visualization pits two primary terms against each other. Then, the use frequency of words that follow these two terms are analyzed. For example, ”war memorial” occurs 531,205 times, while ”peace memorial” occurs only 25,699. A position for each word is generated by looking at the ratio of the two frequencies. If they are equal, the word is placed in the middle of the scale. However, if there is a imbalance in the uses, the word is drawn towards the more frequently related term. This process is repeated for thousands of other word combinations, creating a spectrum of word associations. Font size is based on a inverse power function (uniquely set for each visualization, so you can’t compare across pieces). Vertical positioning is random.”

[<http://www.chrisharrison.net/index.php/Visualizations/WordSpectrum>]

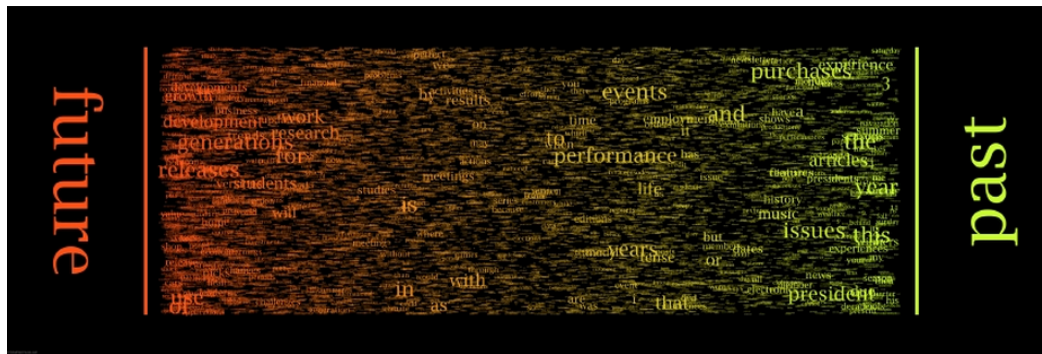


Figure 0.20: ▷ Word Spectrum: Visualizing Google’s Bi-Gram Data by Chris Harrison (2008)

21▷ Word Associations Visualizing Google's Bi-Gram Data by Chris Harrison (2008)

- data: The huge Google bi-gram dataset [Large dataset]
- method: Words rays
- description: “words are bucketed into one of 25 different rays. Each of these represent a different tendency of use (ranging from 0 to 100% in 4% intervals). Words are sorted by decreasing frequency within each ray. I render as many words as can fit onto the canvas. There is a nice visual analogy at play - the ”lean” of each ray represents the strength of the tendency towards one of the two terms. As in the word spectrum visualization, font size is based on a inverse power function (uniquely set for each visualization, so you can’t compare across pieces). Common words (a, the, for, as, etc.) are not shown.”

[<http://www.chrisharrison.net/index.php/Visualizations/WordAssociations>]

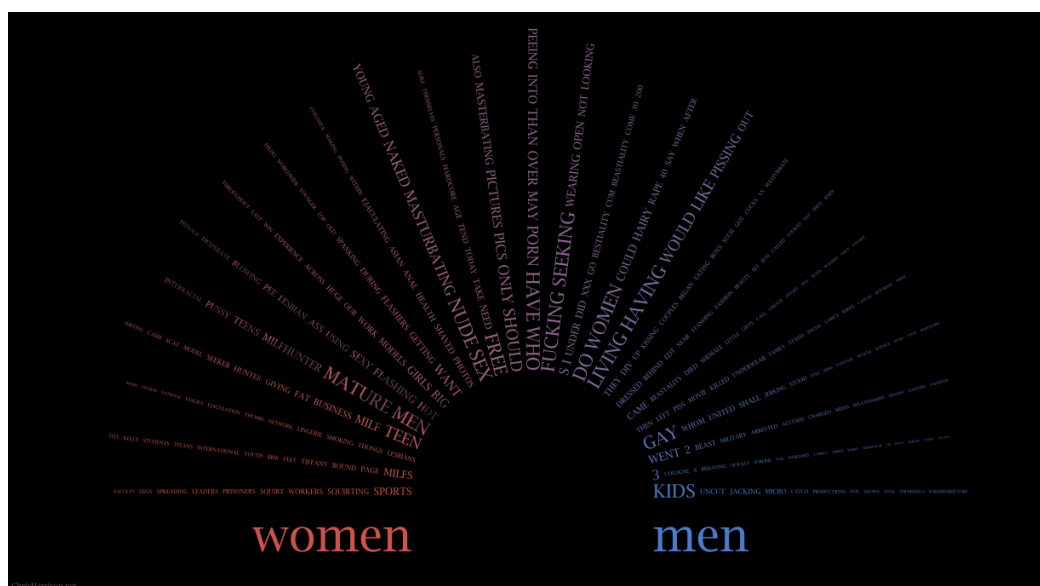


Figure 0.21: Word Associations Visualizing Google's Bi-Gram Data by Chris Harrison (2008)

22▷ Document Arc Diagrams by Jeff Clark (2007)

- data: any text
- method: Arc diagram
- description: Document Arc Diagrams illustrate the similarity structure within a text document by drawing arcs connecting segments of a document that share similar vocabulary.

▷ [<http://www.neoformix.com/2007/DocumentArcDiagrams.html>]

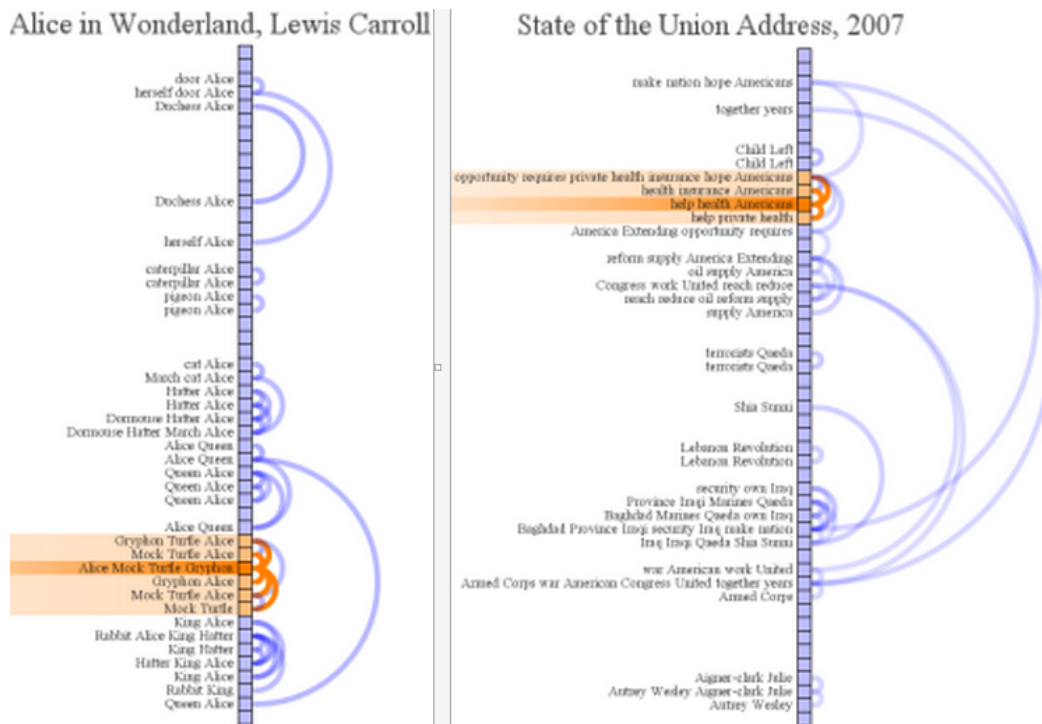


Figure 0.22: Jeff Clark version of arc diagrams

23▷ Gist icons by P. DeCamp, A. Frid-Jimenez, J. Guinness, D. Roy (2005)

- data: any body of literature: 1) The complete set of approximately 7 million USPTO patents; 2) Enron email data set comprised of 500,000 emails; 3) A collection of computer generated speech transcripts from MIT Media Lab Symposia.
- method: interactive radial histogram
- description: The shape contains the semantic profile of a single document where the peaks and valleys are defined by the relatedness of words or concepts to that document.
- paper: *Decamp P., Frid-Jimenez A., Guinness J., Roy D.: Gist icons: Seeing meaning in large bodies of literature. In Proc. of IEEE Symp. on Information Visualization, Poster Session (Oct. 2005).*

▷ [<http://media.mit.edu/cogmac/publications/IEEEIcons.pdf>]

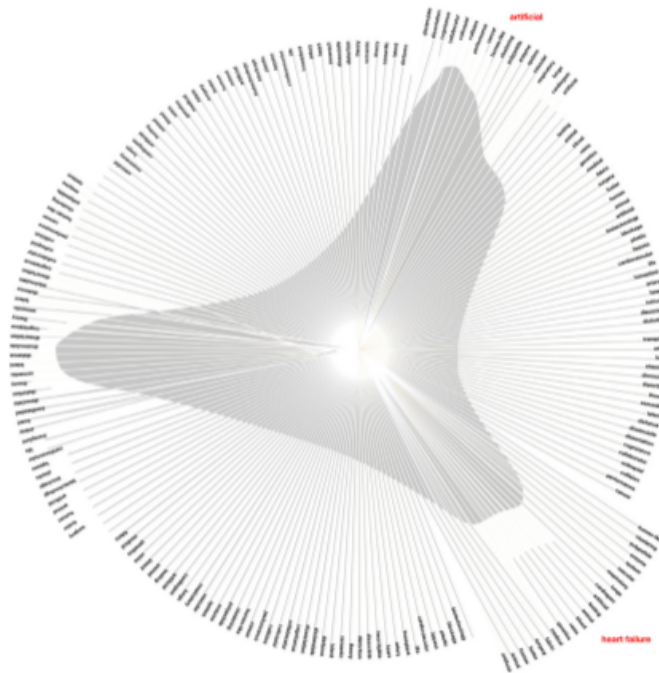
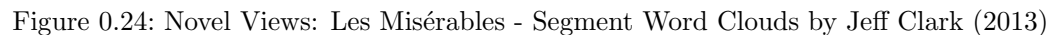


Figure 0.23: Gist icons by P. DeCamp, A. Frid-Jimenez, J. Guinness, D. Roy (2005)

24▷ Novel Views: Les Misérables - Segment Word Clouds by Jeff Clark (2013)

- [<http://neoformix.com/2013/NovelViews.html>]



25▷ Grimm's Fairy Tale Network by Jeff Clark (2013)

- data: 62 stories of the Grimms's Fairy Tales [Small dataset]
- method: 2-dimension networks
- description: "the graphic below is a simple network showing which stories are connected through the use of a common vocabulary. There are three different strengths of connection shown and I've tried to minimize the usual 'hairball' nature of these types of diagrams by only showing the top three connections for a story."

[<http://neoformix.com/2013/GrimmNetwork.html>]

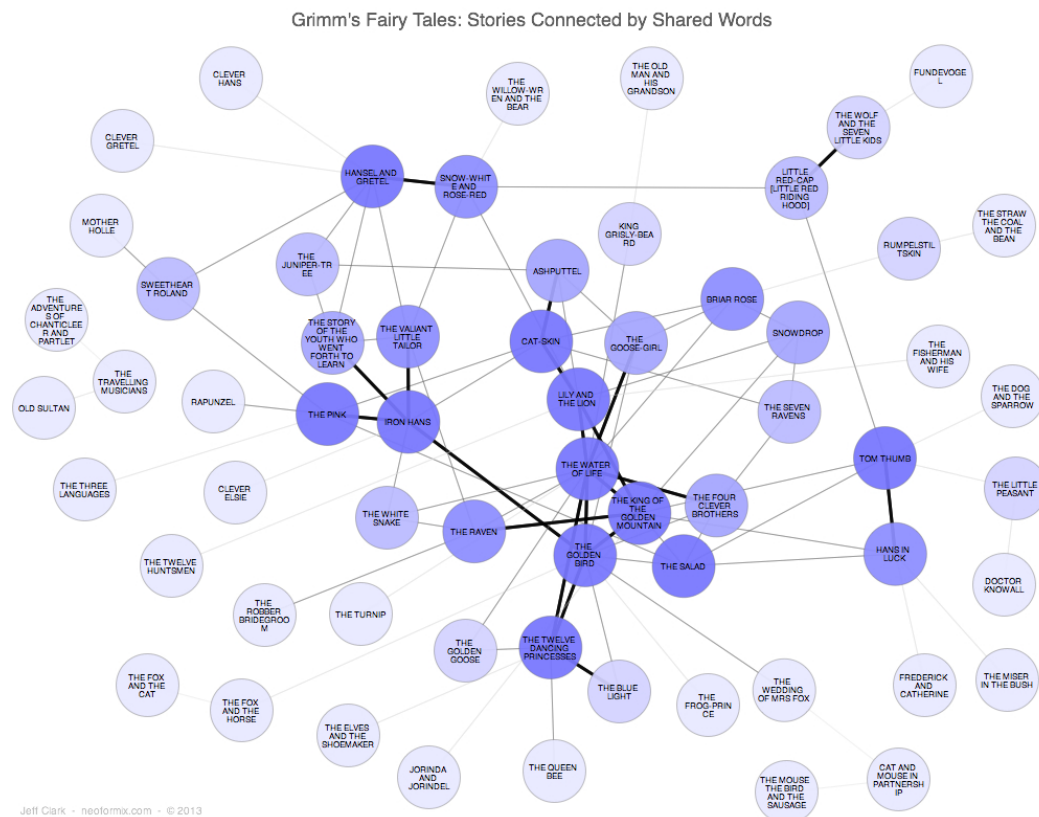


Figure 0.25: Grimm's Fairy Tale Network by Jeff Clark (2013)

26▷ Spot by Jeff Clark (2012)

- data: Twitter [Small dataset]
- method: Multi visualization of tweets based on a search query.
- description: "Spot is an interactive real-time Twitter visualization that uses a particle metaphor to represent tweets. The tweet particles are called spots and get organized in various configurations to illustrate information about the topic of interest."

[<http://neoformix.com/2012/IntroducingSpot.html>]

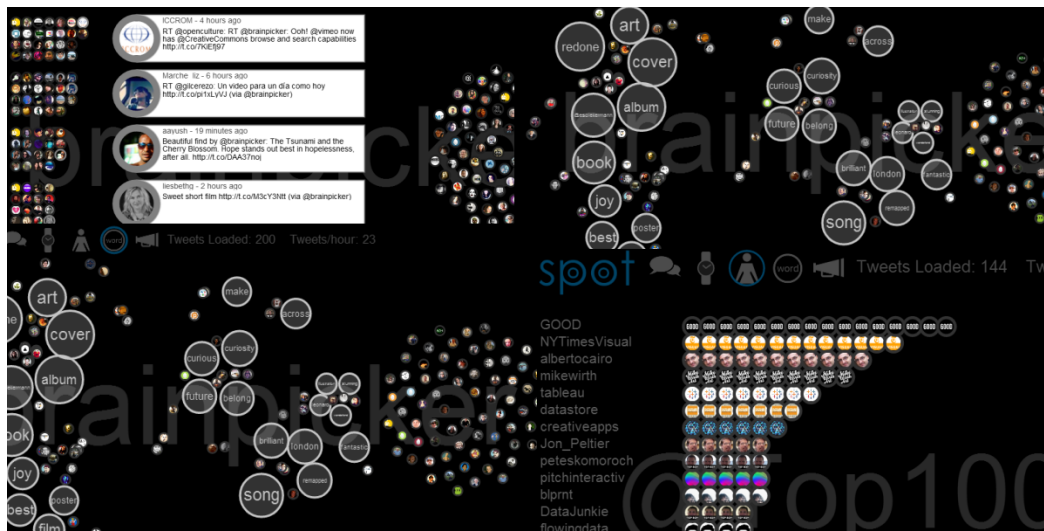


Figure 0.26: Spot by Jeff Clark (2012)

27 ▷ Word storm by Quim Castella and Charles Sutton (2012)

- data: Papers from ICML - International Conference on Machine Learning June 26-July 1, 2012 Edinburgh, Scotland [Small dataset]
- method: Word cloud with fixed word position
- description: visualization of the conference session in the form of a word storm, which is a group of word clouds. The clouds are arranged so that if the same word appears in two clouds, it is in the same position. This is intended to make it easier to see the difference between clouds.
- paper: *Castella, Q., & Sutton, C. (2013). Word Storms: Multiples of Word Clouds for Visual Comparison of Documents. arXiv preprint arXiv:1301.0503.*

[<http://icml.cc/2012/whatson-all/>]



Figure 0.27: Word storm by Quim Castella and Charles Sutton (2012)

28▷ Topic Networks in Proust - Topology by Elijah Meeks, Jeff Drouin (2011)

- data: Marcel Proust texts [Large dataset]
- method: Topic model network
- description: “this is document-topic network representation. This visualization shows the relationships abd topics of a collection of documents. Visually this is represented by the distances between documents, topics and documents-topic.”

[<https://dhs.stanford.edu/algorithmic-literacy/topic-networks-in-proust/>]

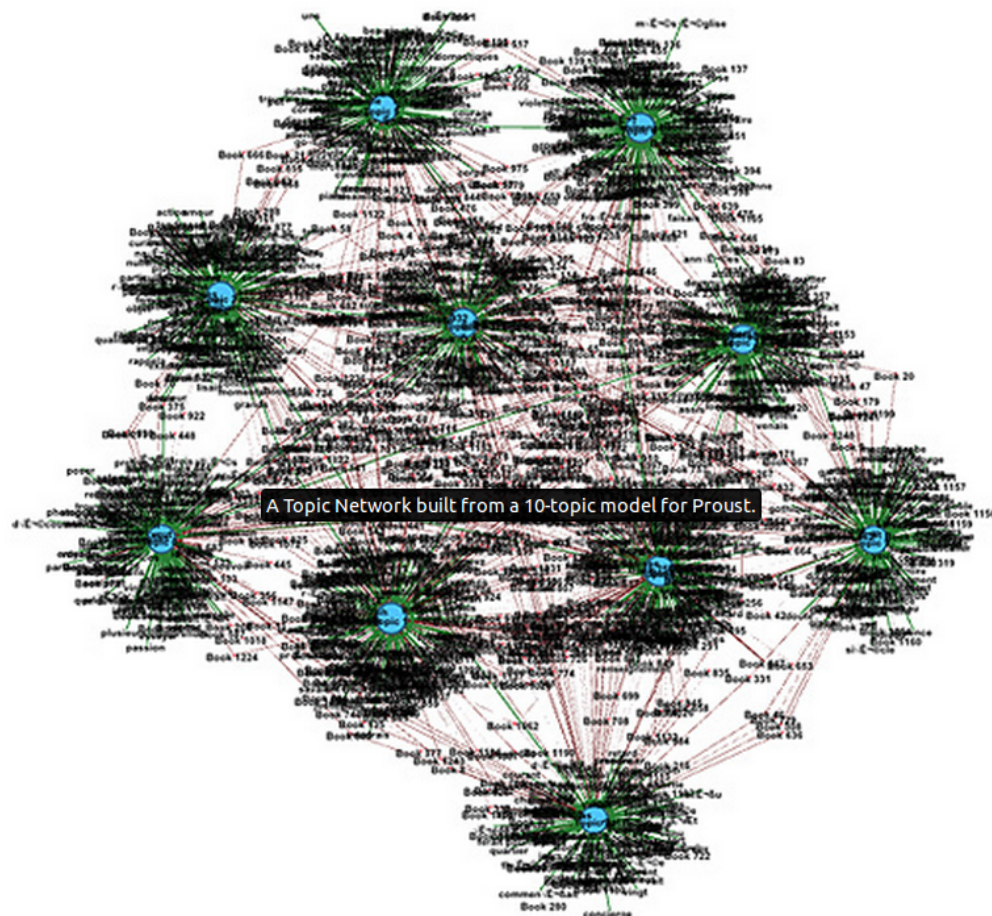


Figure 0.28: Topic Networks in Proust - Topology by Elijah Meeks, Jeff Drouin (2011)

29▷ Notabilia. 100 Longest Article for Deletion [AfD] discussions on Wikipedia by Dario Taraborelli, Giovanni Luca Ciampaglia (data and analysis) and Moritz Stefaner (visualization). (2010)

- data: 100 Longest Article for Deletion [AfD] of Wikipedia [Small dataset]
- method: Original. Bouquet of edition lines
- description: “visualization of debate deletion of entries in Wikipedia. Each time a user joins an AfD discussion and recommends to keep, merge, or redirect the article a green segment leaning towards the left is added. Each time a user recommends to delete the article a red segment leaning towards the right is added. As the discussion progresses, the length of the segments as well as the angle slowly decay.”

[<http://notabilia.net/>]



Figure 0.29: Notabilia. 100 Longest Article for Deletion [AfD] discussions on Wikipedia by Dario Taraborelli, Giovanni Luca Ciampaglia (data and analysis) and Moritz Stefaner (visualization). (2010)

30▷ X by Y by Moritz Stefaner (2009)

- data: Almost 40.000 submissions to the Prix Ars Electronica, from the early beginnings in 1987 up to 2009 [Large dataset]
- method: Visual spots in groups and colors.
- description: “X by Y visualizes all submissions to the Prix Ars Electronica, from the early beginnings in 1987 up to 2009. The goal is to characterize the ”ars world” in quantitative terms. A series of diagrams groups and juxtaposes the submissions by years, categories, prizes and countries. The graphics are composed of little dots (each representing a single submission) to provide a visual scale for the statistical statements and thematize the relation of the totality and the individual.”
- Book reference: *Stefaner, M., Ferré, S., Perugini, S., Koren, J., & Zhang, Y. (2009). User interface design. In Dynamic Taxonomies and Faceted Search (pp. 75-112). Springer Berlin Heidelberg.*

[<http://moritz.stefaner.eu/projects/x-by-y/>]

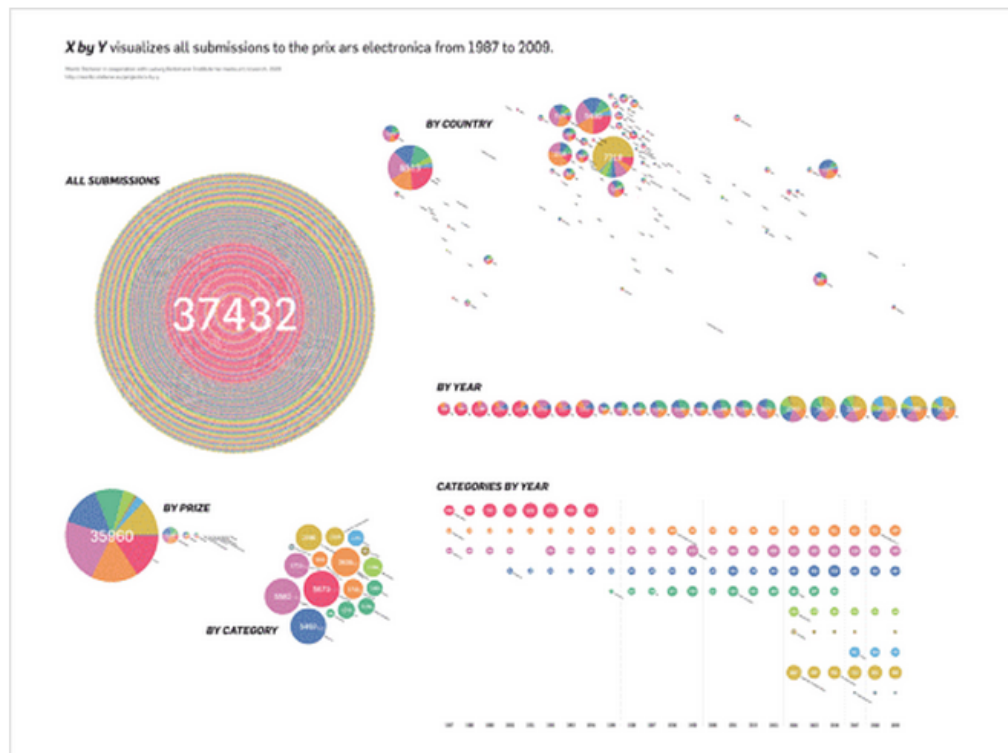


Figure 0.30: X by Y by Moritz Stefaner (2009)

31▷ Search Clock by Chris Harrison (2008)

- data: Search engine queries [Large dataset]
- method: Clock with queries
- description: "I was curious to see if data from search engines would support my anecdotal observations. I built a simple clock-like visualization that displays the top search terms over a 24-hour period. Displaying search terms in a cyclical layout (like a clock) allows continuous examination of trends that would otherwise be broken up. The data I had access to was both large and noisy. In response, I combined hourly data into week or year averages."

[<http://www.chrisharrison.net/index.php/Visualizations/SearchClock>]

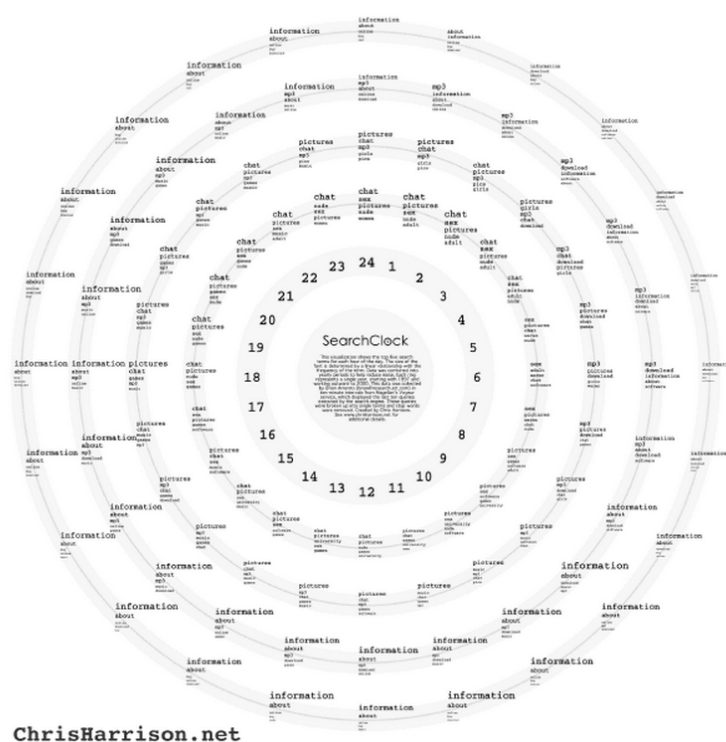


Figure 0.31: Search Clock by Chris Harrison (2008)

32▷ Digg Rings by Chris Harrison (2008)

- data: Digg data - top 10 most digg stories of the day for May 24, 2007 to May 23, 2008 [Large dataset]
- method: Tree-ring-like visualizations
- description: “using the Digg API, I grabbed the top 10 most-dugg stories of the day (by midnight) for the past year - May 24, 2007 to May 23, 2008. I then rendered a series of tree-ring-like visualizations (moving outwards in time). Rings are colored according to Digg’s eight top-level categorizations (see key at bottom of page). Ring thickness is linearly proportional to the number of diggs the story received. I also made a pair of visualizations using Digg’s entire archive, which goes back to December 1, 2004.”

[<http://www.chrisharrison.net/index.php/Visualizations/DiggRings>]

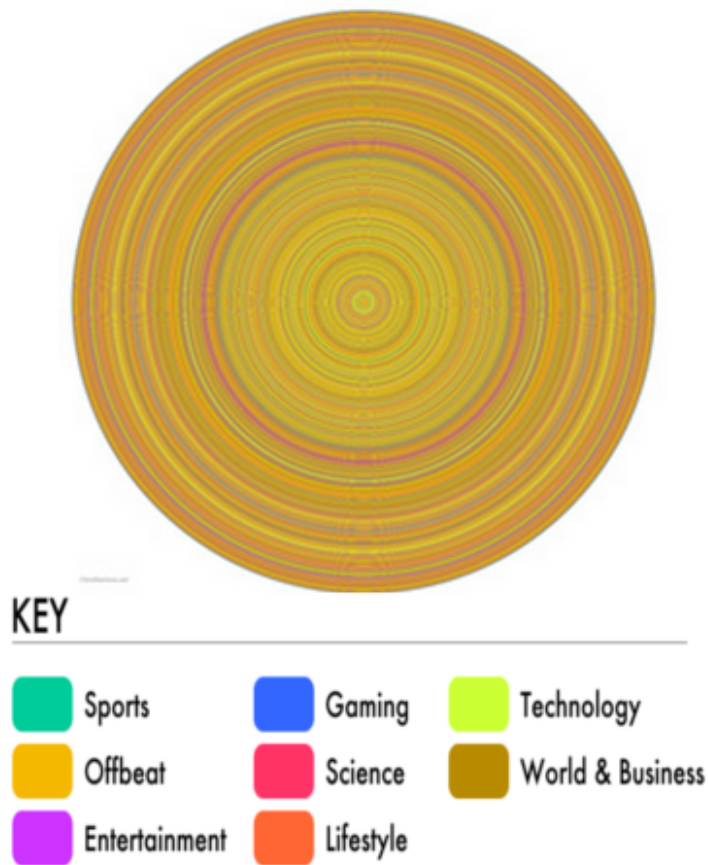


Figure 0.32: Digg Rings by Chris Harrison (2008)

33▷ Royal Society Archive by Chris Harrison (2008)

- data: titles of papers of the Royal Society Archive (1665-2005) [Large dataset]
- method: time-line
- description: "The Royal Society recently provided access to an archive of papers published in the scientific academy's prestigious journals. Some 25 thousand scholarly works are represented, which date from 1665 to 2005. Many notable scientific minds are represented, including Isaac Newton, Michael Faraday and Charles Darwin. This interesting data set was ripe for some visual tinkering. The database I used was put together by Brian Amento and Mike Yang of AT&T Labs."

[<http://www.chrisharrison.net/index.php/Visualizations/RoyalSociety>]



Figure 0.33: Detail of the visualization Royal Society Archive by Chris Harrison (2008)

34▷ WikiViz: Visualizing Wikipedia by Chris Harrison (2007)

- data: Wikipedia [Large dataset]
- method: network
- description: "Wikipedia is an interesting dataset for visualization. As an encyclopedia, it's articles span millions of topics. Being a human edited entity, connections between topics are diverse, interesting, and sometimes perplexing - five hops takes you from subatomic particles to Snoop Dog. Wikipedia is revealing in how humans organize data and how interconnected seemingly unrelated topics can be."

[<http://www.chrisharrison.net/index.php/Visualizations/WikiViz>]

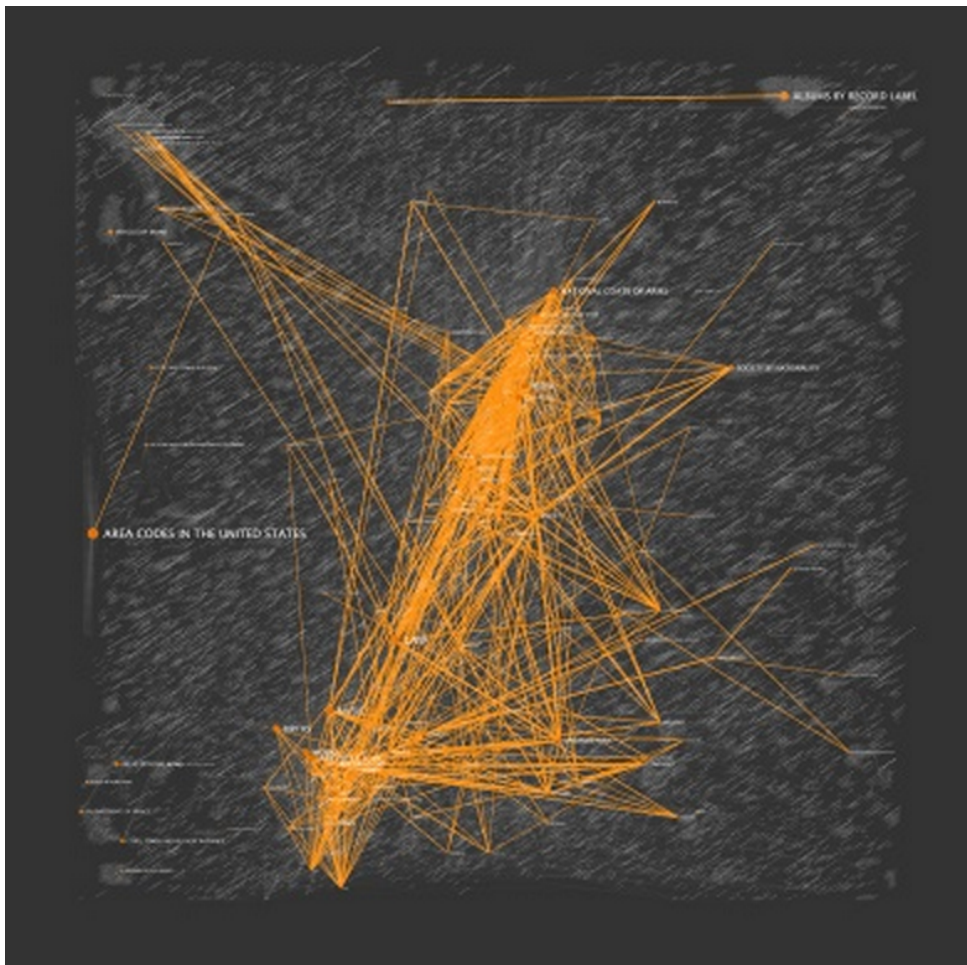


Figure 0.34: WikiViz: Visualizing Wikipedia by Chris Harrison (2007)

35▷ Area by Jaume Nualart (2007)

- data: Any collection of texts with parameterizable meta-data to discrete values [Small dataset]
- method: Database representation and browsing tool.
- description: AREA represents the whole dataset, it gives a size-overview of the data, is a data browser, is a data visualization, is a whole data understanding, is interactive, data can be filtered, some time using Area teaches about the main characteristics of the represented data.

[<http://nualart.com/area2>]



Figure 0.35: Area by Jaume Nualart (2007)

36 ▸ Visualizing Activity on Wikipedia with Chromograms by M. Wattenberg, F.B. Viégas, and K. Hollenbach (2004)

- data: Wikipedia [Large dataset]
- method: Chromograms
- description: a chromogram describes the unique edit pattern of a Wikipedian over time by categorizing and color coding edits. Chromograms for distinct persons (and bots!) can be markedly dissimilar.
- Wattenberg, M., Viégas, F. B., & Hollenbach, K. (2007). *Visualizing activity on Wikipedia with chromograms*. In *Human-Computer Interaction-INTERACT 2007* (pp. 272-287). Springer Berlin Heidelberg.

[http://pensivepuffin.com/dwmcphd/syllabi/info447_wi12/readings/wk09-Organizing/wattenberg.Chromogram.INTERACT08.pdf]

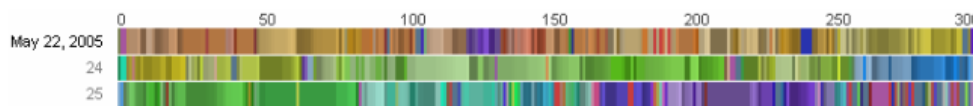


Figure 0.36: Visualizing Activity on Wikipedia with Chromograms by M. Wattenberg, F.B. Viégas, and K. Hollenbach (2004)

37▷ Kartoo/Ujiko by Laurent Baleyrier and Nicholas Baleyrier (2001)

- data: Internet web pages [Small dataset]
- method: Search results as a map of web pages with tagged edges
- description: “Kartoo and Ujiko have been the more advanced search engine interfaces for a wide use. The results were presented as a map of web pages with tagged edges. Kartoo (later Ujiko) was born in 2001 and shut down in 2010.”

[<http://en.Wikipedia.org/wiki/Kartoo>]

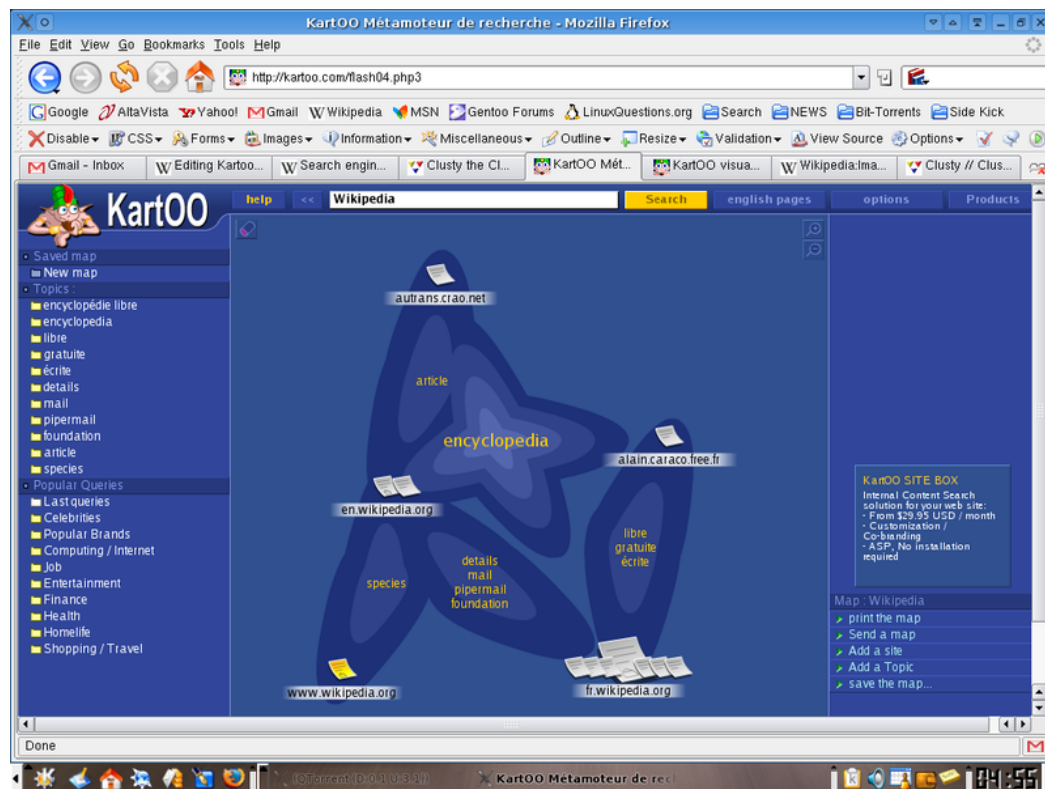


Figure 0.37: Kartoo/Ujiko by Laurent Baleyrier and Nicholas Baleyrier (2001)

38▷ Touchgraph by TouchGraph, LLC. (2001)

- data: Search engine results [Large dataset]
- method: Network visualization of search results
- description: TouchGraph was founded in 2001 with the creation of the original visual browser for Google. Since then, millions of people have used TouchGraph's tools to discover the relationships contained in Google, Amazon, Wikis, and other popular information sources.

[<http://www.touchgraph.com/seo>]

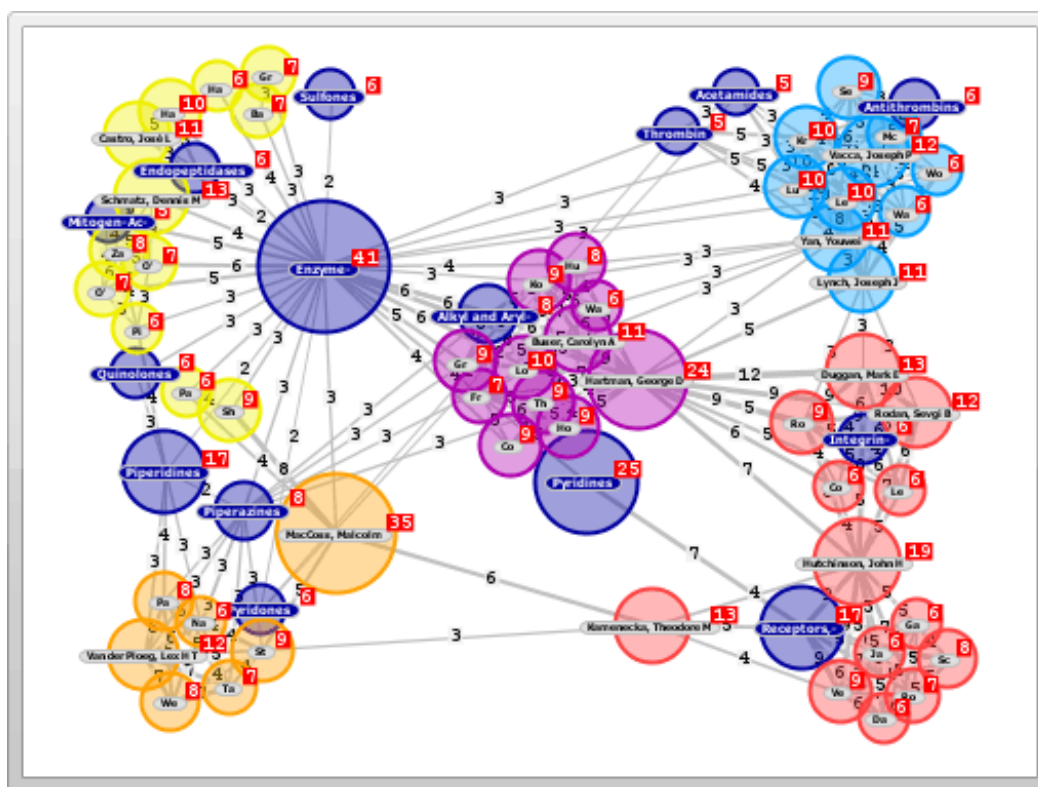


Figure 0.38: Touchgraph by TouchGraph, LLC. (2001)

39▷ HotSauce by Ramanathan V. Guha (1996)

- data: Websites relationships written in Meta Content Framework (MCF), a precursor of RDF. [Small dataset]
- method: 3D interactive network
- description: Ramanathan V. Guha: “HotSauce worked as a plug-in to an existing browser so that when a hyperlink to a MCF-enabled website was selected the user was dropped into a first-person perspective view of the Web. It was a videogame view with Web pages floating as brightly colored blocks in an infinite black space, something like the view from a starship cockpit navigating through some strange asteroid field. It was easy to fly into and around the space, using the mouse to guide the direction of flight and holding down buttons to go forwards and backwards. A page could be accessed by simply double-clicking on the relevant block. A 3D immersive environment where I could move around an Antarcti.ca like space, moving things around, interacting with others in that space. And I should be able to do this wherever I am!”

[http://mappa.mundi.net/maps/maps_018/]

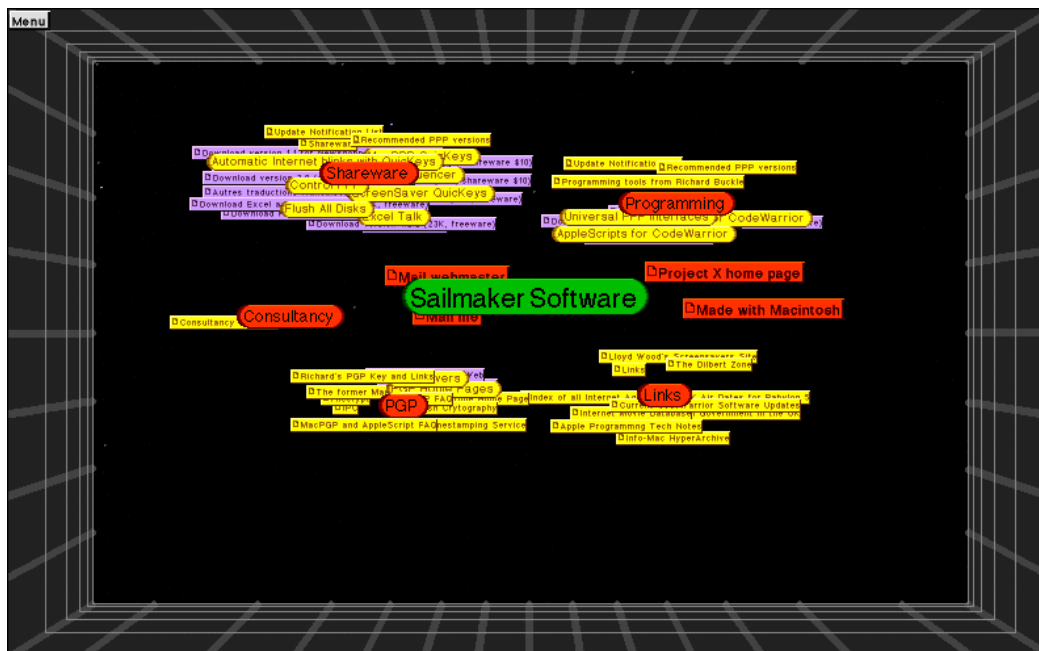


Figure 0.39: HotSauce by Ramanathan V. Guha (1996)

Collections of aggregations visualization

40 ▷ Grimm's Fairy Tale Metrics by Jeff Clark (2013)

- data: 62 stories of the Grimms's Fairy Tales [Small dataset]
- method: Sortable matrix with links to the dataset
- description: very complete metric analysis of the 62 stories of the Grimms brothers. This is a high quality example of a tool for linguistics analysis.

[<http://neoformix.com/2013/GrimmStoryMetrics.html>]

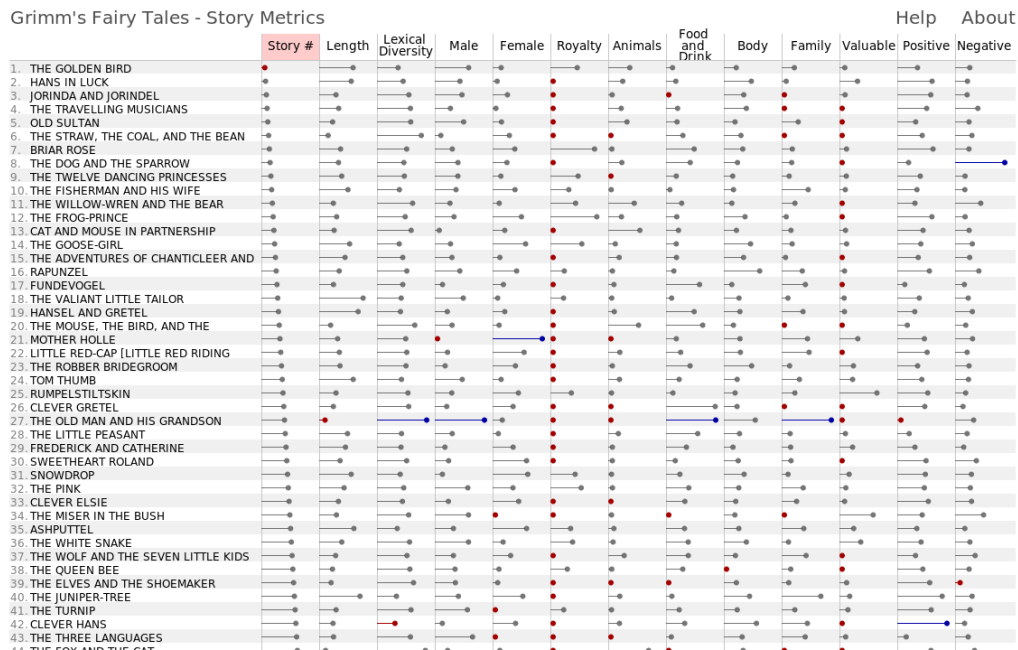


Figure 0.40: Grimm's Fairy Tale Metrics by Jeff Clark (2013)

41▷ Termite: Visualization Techniques for Assessing Textual Topic Models by Jason Chuang, Christopher D. Manning, Jeffrey Heer (2012)

- data: any text corpora [Large dataset]
- method: Matrix/tabular view
- description: it is matrix view to support the assessment of topical term distributions and enable the comparison of latent topics. And, in general, it is a visualization of topic models and their terms distribution.
- paper: Chuang, J., Manning, C. D., & Heer, J. (2012, May). *Termite: Visualization techniques for assessing textual topic models*. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (pp. 74-77). ACM.

[<http://vis.stanford.edu/papers/termite>]

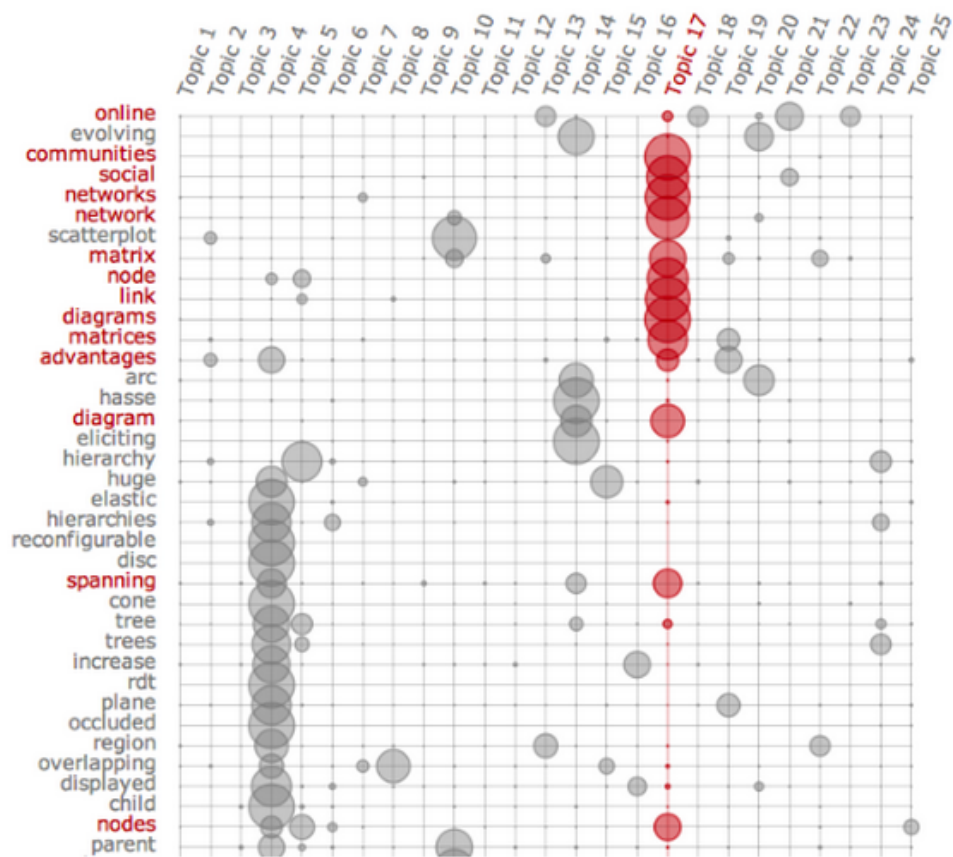


Figure 0.41: Termite: Visualization Techniques for Assessing Textual Topic Models by Jason Chuang, Christopher D. Manning, Jeffrey Heer (2012)

42▷ Pediameter by Müller-Birn, Benedix and Hantke (2011)

- data: Wikipedia [N/A]
- method: time-line + arduino indicator
- description: a live visualization of Wikipedia Edits using pixel block in a time-line and an arduino indicator

[<http://l3q.de/pediameter/>]

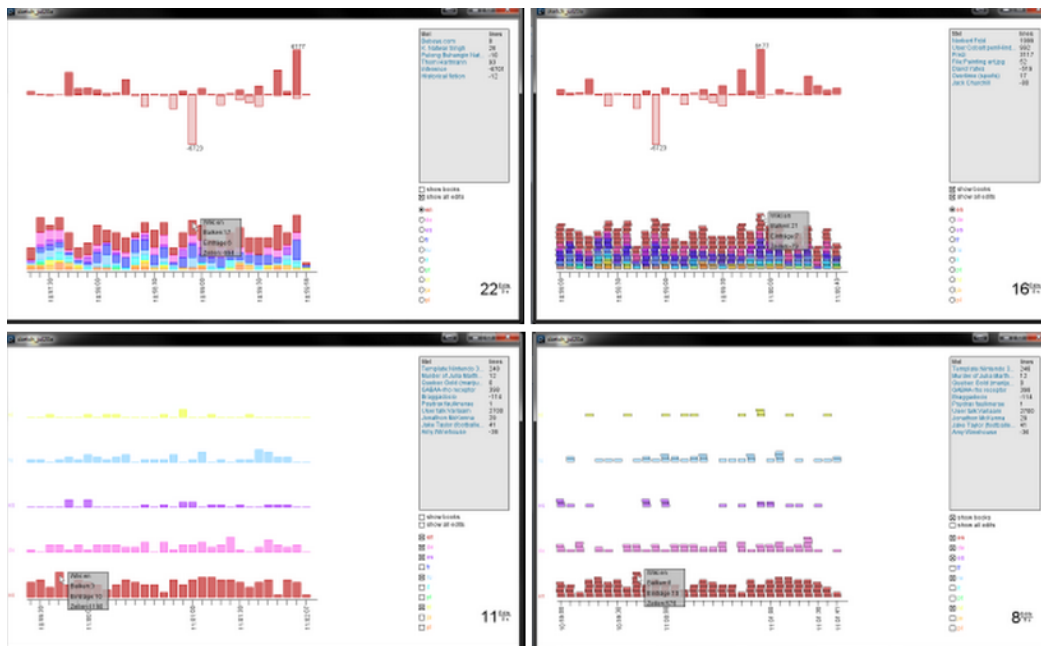


Figure 0.42: Pediameter by Müller-Birn, Benedix and Hantke (2011)

43 ▸ Web Seer by Fernanda Viégas & Martin Wattenberg (2009)

- data: Google search suggestions [Large dataset]
- method: Connected trees
- description: a visualization that lets you compare two Google suggest queries.

[<http://hint.fm/seer/>]



Figure 0.43: Web Seer by Fernanda Viégas & Martin Wattenberg (2009)

44 ▸ Web Trigrams: Visualizing Google's Tri-Gram Data by Chris Harrison (2008)

- data: Google n-gram dataset (2006) [large dataset]
- method: specific. Connected trees
- description: "As soon as I got my hands on the data, I quickly got to work on some straight forward visualizations. The first type compares two sets of trigrams, each starting with a different word. One visualization compares 'He' with 'She', while the other uses 'I' and 'You'. In the case of the 'He' vs. 'She', the top 120 trigrams for each were identified. The frequencies of the second word in the trigrams were combined and sorted, and rendered in decreasing frequency-of-use order. A similar process was used to create a ranking for the third (and final) word in the trigrams. Words are sized according to the square root of their use frequencies. The color-coded lines act like paths (a tree structure), enumerating all of the trigrams. The process was identical for the 'I' and 'You' version, except that only the top 75 trigrams were used."

[<http://www.chrisharrison.net/index.php/Visualizations/WebTrigrams>]

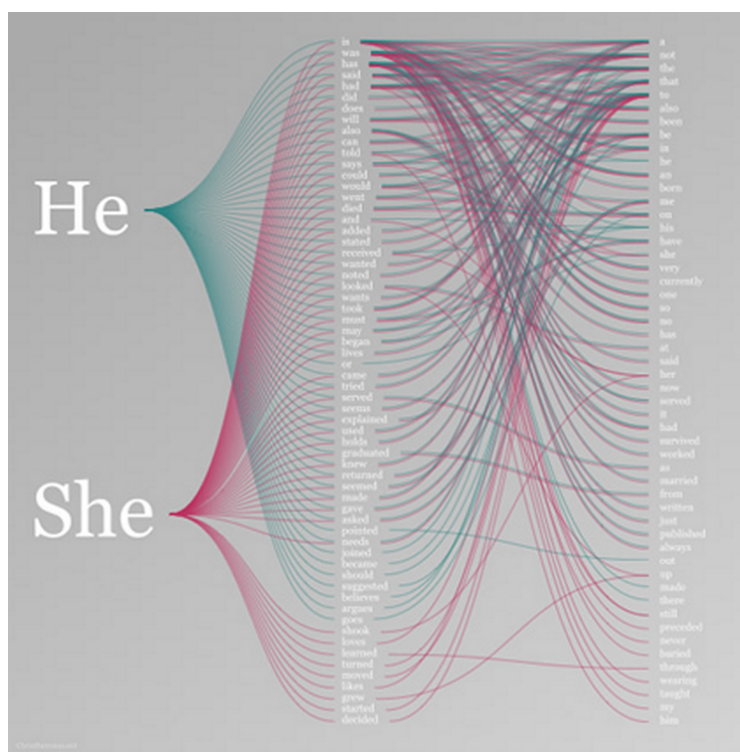


Figure 0.44: Differences in how the he and she subjects are used.

45▷ FeaturedLens by A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. (2007)

- data: President Bush's eight annual speeches (2001-2007)
- method: rich interface for search, visualize and explore texts
- description: FeatureLens visualizes a text collection at several levels of granularity and enables users to explore interesting text patterns. The current implementation focuses on frequent itemsets of n-grams, as they capture the repetition of exact or similar expressions in the collection. Users can find meaningful co-occurrences of text patterns by visualizing them within and across documents in the collection. This also permits users to identify the temporal evolution of usage such as increasing, decreasing or sudden appearance of text patterns. The interface could be used to explore other text features as well.
- paper: *Don A., Zheleva E., Gregory M., Tarkan S., Auvil L., Clement T., Shneiderman B., Plaisant C.: Discovering interesting usage patterns in text collections: Integrating text mining with visualization. In Proc. of the Conf. on Information and Knowledge Management (2007).*

[<http://hcil2.cs.umd.edu/trs/2007-08/2007-08.pdf>]

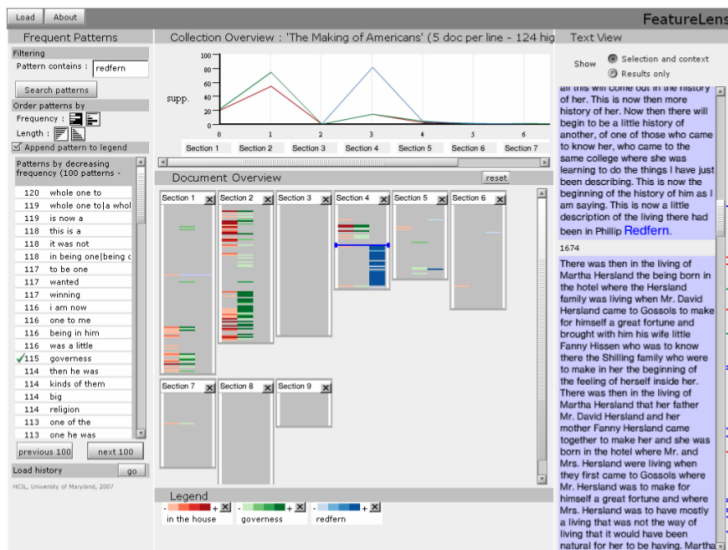


Figure 0.45: Screenshot of the FeatureLens interface.

46 ▷ **Newsmap by Marcos Weskamp (2004)**

- data: Google news [Small dataset]
- method: Treemap
- description: "Newsmap is an application that visually reflects the constantly changing landscape of the Google News news aggregator. Google News automatically groups news stories with similar content and places them based on algorithmic results into clusters. In Newsmap, the size of each cell is determined by the amount of related articles that exist inside each news cluster that the Google News Aggregator presents. In that way users can quickly identify which news stories have been given the most coverage, viewing the map by region, topic or time"

[<http://marumushi.com/projects/newsmap>]



Figure 0.46: Newsmap by Marcos Weskamp (2004)

47▷ TheMail by Fernanda B. Viégas, Scott Golder, Judith Donath (2006)

- data: Collection of emails conversation. [Small dataset]
- method: Speciphic interface. Time line with columns of words
- description: "a visualization that portrays relationships using the interaction histories preserved in email archives. Using the content of exchanged messages, it shows the words that characterize ones correspondence with an individual and how they change over the period of the relationship."
- paper: Viégas, F. B., Golder, S., & Donath, J. (2006, April). *Visualizing email content: portraying relationships from conversational histories*. In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (pp. 979-988). ACM.

[http://smg.media.mit.edu/papers/Viegas/themail/viegas_themail.pdf]

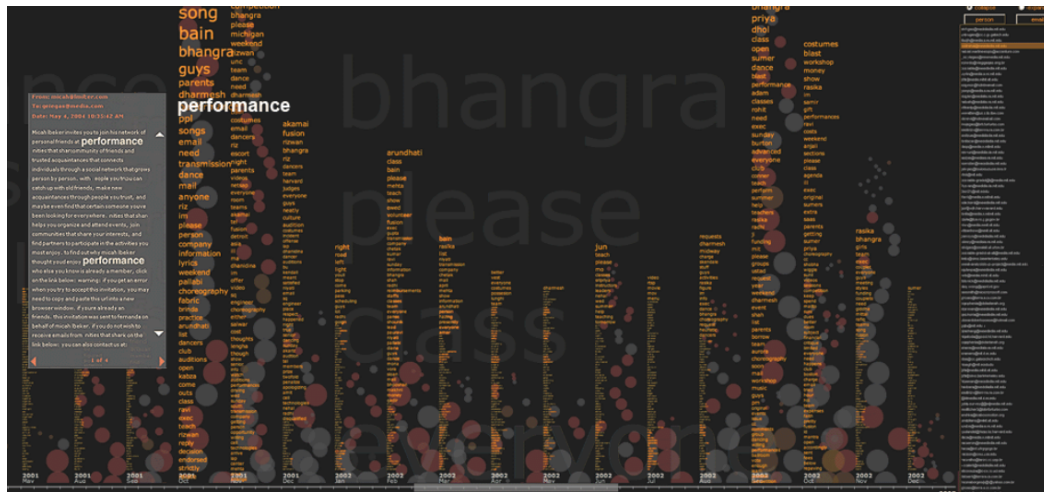


Figure 0.47: TheMail by Fernanda B. Viégas, Scott Golder, Judith Donath (2006)

48▷ WebBook by Stuart K. Card, George G. Robertson, and William York (1996)

- data: Search engine results [Small dataset]
- method: 3D interactive book of HTML pages
- description: WebBook was a dynamic multimedia contents constructor from lists of Internet pages. It is very inspiring the reading of the 1996 paper to understand the history of the world wide web.
- paper: Card, S. K., Robertson, G. G., & York, W. (1996, April). *The WebBook and the Web Forager: an information workspace for the World-Wide Web*. In *Proceedings of the SIGCHI conference on Human factors in computing systems: common ground* (pp. 111-ff). ACM.

[<http://www.sigchi.org/chi96/proceedings/papers/Card/skc1txt.html>]



Figure 0.48: WebBook by Stuart K. Card, George G. Robertson, and William York (1996)

49▷ Dotplot Applications by Jonathan Helfman (1994)

- data: Any text [Large dataset]
- method: Dotplot
- description: a bit minimalistic and very effective method to check similarity in large amount of texts, including multilanguage texts or computer code.
- paper: *Helfman, J. (1996). Dotplot patterns: a literal look at pattern languages. Theory and Practice of Object Systems, 2(1), 31-41.*

[<http://imagebeat.com/index.php?id=17>]

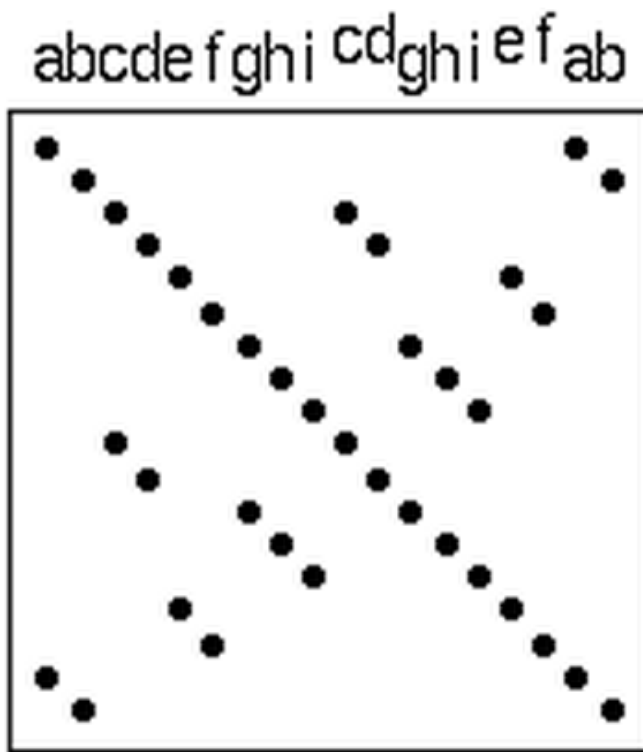


Figure 0.49: Dotplot Applications by Jonathan Helfman (1994)